Nama: Bernard Hugo

NIM: 2540124450

Kelas: LA05

UAS Data Mining & Visualization

Odd Semester

Link video: https://youtu.be/o_6JKsIv8Nc

1. The insurance.xlsx dataset contains 438 observations (rows) and 7 features (columns). The dataset contains 3 numerical predictors (age, bmi, and children) and 3 nominal predictors (sex, smoker and region). The purposes of this exercise to look into different features to observe their relationship, and plot a regression model based on several features of individual such as age, physical/family condition and location against their existing medical expense to be used for predicting future medical expenses of individuals that help medical insurance to make decision on charging the premium. Please do analysis using R with some tasks as follows:

a. Define the most suitable regression model based on all predictor variables and response variable. Explain your answer.

Jawab:

Regression model yang sangat cocok berdasarkan variable-variabel predictor/independen dan variable response/dependen adalah linear regression. Linear regression sangat berguna dalam mencocokkan model regresi yang menggambarkan hubungan antara variable-variabel prediktor dengan sebuah variabel respons numerik.

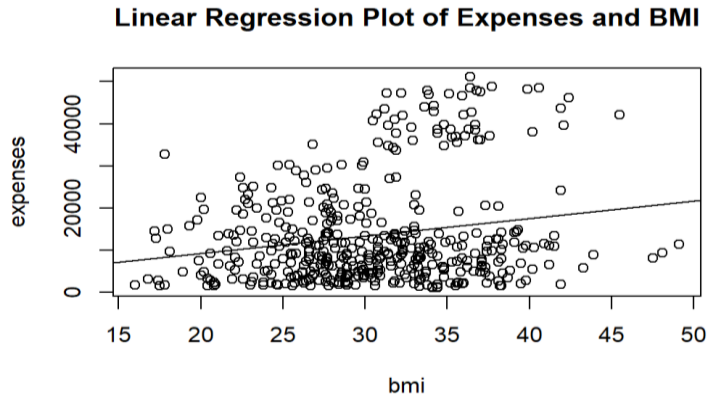
Dari dataset insurance.xlsx, variable response numeriknya adalah expenses yaitu besarnya pengeluaran seseorang untuk asuransi atau jaminan kesehatan. Variabel-variabel predictor seperti: children, age, bmi, region, sex, dan smoker memiliki hubungan dengan expenses yang akan mempengaruhi besarnya biaya pengeluaran untuk biaya pengobatan dan lain-lain.

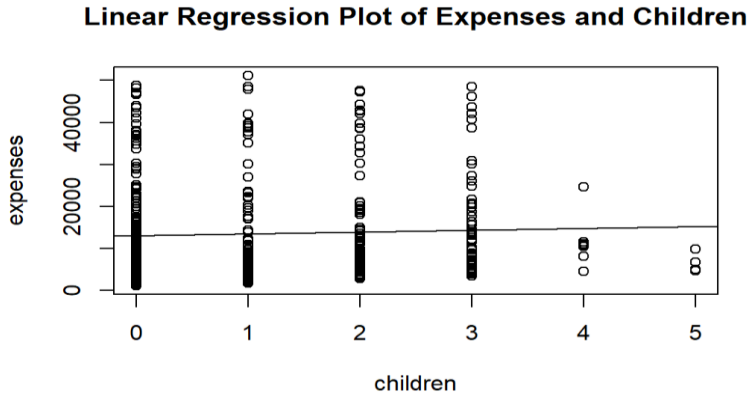
b. Construct the model regression you choose based on answer (a).

Jawab:

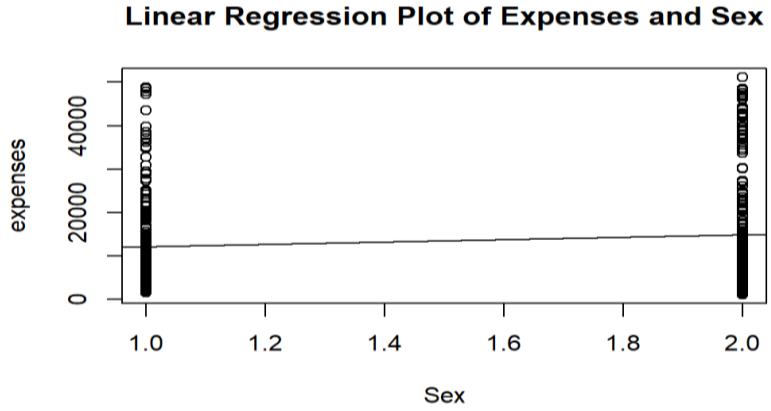
```
library(readxl)

df1 <- read_excel("Insurance.xlsx")
df1

summary(df1)

model1 <- lm(expenses ~ ., data = df1)
model1

summary(model1)
```

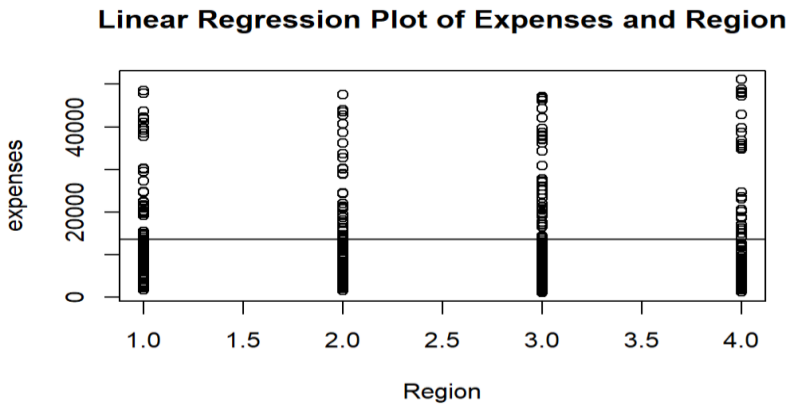
Hasil plot:

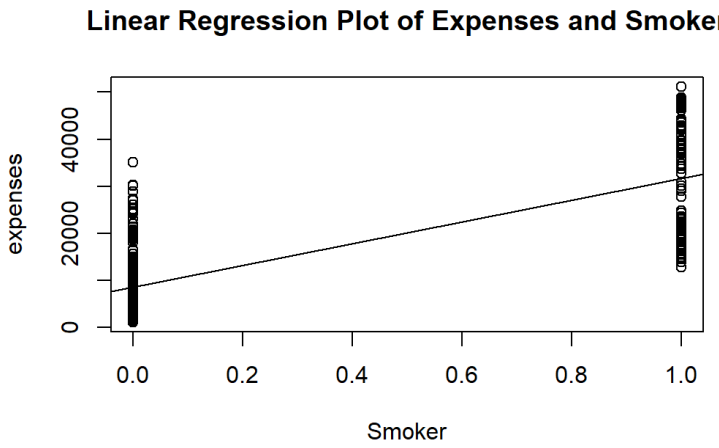





Linear Regression Plot of Expenses and Age
expenses
0
20000
40000
20
30
40
50
60
age


Linear Regression Plot of Expenses and Sex
expenses
0
20000
40000
1.0
1.2
1.4
1.6
1.8
2.0
Sex


Linear Regression Plot of Expenses and Region
expenses
0
20000
40000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Region


Linear Regression Plot of Expenses and Smoker
expenses
0
20000
40000
0.0
0.2
0.4
0.6
0.8
1.0
Smoker

c. Interpret your regression model. Give the details.

Jawab:

Call:

lm(formula = expenses ~ ., data = df1)

Residuals:

Min 1Q Median 3Q Max

-12191 -3104 -1018 1587 24801

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13470.37	1776.81	-7.581	2.14e-13
age	268.36	20.27	13.236	< 2e-16
sexmale	935.59	583.78	1.603	0.110
bmi	346.87	51.50	6.736	5.27e-11
children	348.36	240.25	1.450	0.148
smokeryes	23250.05	708.17	32.831	< 2e-16
regionnorthwest	450.84	827.29	0.545	0.586
regionsoutheast	-736.50	812.05	-0.907	0.365
regionsouthwest	404.34	835.84	0.484	0.629

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

Residual standard error: 6055 on 429 degrees of freedom

Multiple R-squared: 0.7587,

Adjusted R-squared: 0.7542

F-statistic: 168.6 on 8 and 429 DF,

p-value: < 2.2e-16

Dari hasil summary linear regression model dengan variable respons expenses, bisa dijelaskan bahwa linear model tersebut memiliki nilai minimum sebesar -12191. Setelah itu terdapat juga nilai 1 quartile = -3104, median = -1018, 3 quartile = 1587, dan nilai masksimum = 24801. P value pada variable smoker dengan data yes dan variabel age memiliki nilai lebih kecil dari 2e-16.

Selain itu, hasil summary(model1) juga menunjukkan bahwa linear model expenses memiliki residual standard error sebesar 6055 pada 429 derajat kebebasan, Multiple R-squared: 0.7587, Adjusted R-squared: 0.7542, F-statistic: 168.6 pada 8 dan 429 DF, serta p-value lebih kecil dari 2.2e-16.

Dari plot-plot atau grafik-grafik diatas, variable predictor/independen yang memiliki hubungan yang erat dengan expenses adalah bmi, children, smoker, dan region. Bisa dijelaskan bahwa biaya pengeluaran yang tinggi dimiliki oleh seseorang yang memiliki bmi dengan nilai diantara nilai minimum dan maksimumnya. Individu atau keluarga yang memiliki jumlah anak yang banyak akan memiliki biaya pengeluaran yang lebih besar. Karena suatu keluarga harus memenuhi kebutuhan hidup anak-anaknya. Daerah yang ditinggali seseorang juga berpengaruh pada pengeluaran seseorang contohnya jika tinggal di daerah perkotaan yang lebih maju maka biaya pengeluarannya lebih besar.

Sedangkan age dan sex memiliki kaitan yang tidak erat dengan expenses. Karena baik jenis kelamin laki-laki maupun perempuan dan usia berapapun, semua tergantung pada gaya hidup seseorang dalam memenuhi biaya hidupnya masing-masing. Tetapi untuk usia, semakin bertambah usia seseorang maka biaya pengeluarannya menjadi besar terutama dalam biaya pengobatan untuk usia yang sudah tua.

d. Do the significance test for all parameters and analyze the goodness of the model.

Dari hasil summary(df1) didapatkan p-value (pr(> |t|)) yang menjadi hasil dari significance test semua parameter pada dataset. Setelah itu, untuk menganalisis kebagusan model dilihat dari R-squared yang ada pada hasil summary(df1).

	Pr(>|t|)
(Intercept)	2.14e-13
age	< 2e-16
sexmale	0.110
bmi	5.27e-11
children	0.148
smokeryes	< 2e-16
regionnorthwest	0.586
regionsoutheast	0.365
regionsouthwest	0.629

Multiple R-squared: 0.7587

Adjusted R-squared: 0.7542

P – value: < 2.2e-16

Hasil significance test untuk semua parameter bisa dilihat pada p-value setiap variabel dan p-value rata-ratanya adalah lebih kecil dari 2.2e-16.

Untuk analisis kebaikan model, dilihat dari kedua R-squared yang dihasilkan. Maka model regresinya dapat dikatakan hampir bagus karena kedua nilai R-squared yaitu Multiple R-squared = 0.7587 dan Adjusted R-squared = 0.7542 semakin mendekati 1.

2. Breast cancer is the most common cancer amongst women in the world. It accounts for 25% of all cancer cases, and affected over 2.1 Million people in 2015 alone. It starts when cells in the breast begin to grow out of control. These cells usually form tumors that can be seen via X-ray or felt as lumps in the breast area. The key challenges against it's detection is how to classify tumors into malignant (cancerous) or benign(non cancerous). These variables consist of Unique ID, diagnosis Target: M - Malignant B – Benign, Radius of Lobes, Mean of Surface Texture, Outer Perimeter of Lobes, Mean Area of Lobes, Mean of Smoothness Levels, Mean of Compactness, Mean of Concavity, Mean of Cocave Points, Mean of Symmetry, Mean of Fractal Dimension. We ask you to complete the analysis of classifying based on the Breast Cancer Wisconsin (Diagnostic).xlsx Dataset. Please do analysis using R with some tasks as follows:

a. Define the most suitable classification method based on all predictor variables and response variable. Explain your answer.

Jawab:

Metode klasifikasi yang cocok berdasarkan variable-variabel predictor dan variable-variabel response dari dataset "Breast Cancer Wisconsin (Diagnostic).xlsx" adalah klasifikasi logistic regression. Karena logistic regression sangat berguna untuk memprediksi hasil data biner berdasarkan variable-variabel independen. Sesuai dengan datasetnya, variable dependen dari dataset tersebut adalah diagnosis dengan data binernya adalah: "M" – Maligna dan "B" – Benign.

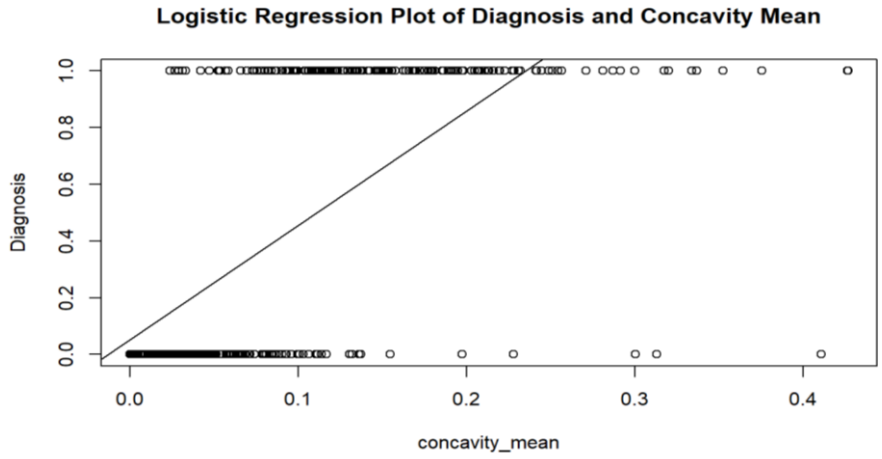
Nilai-nilai pada variabel-variabel independen yang ada pada dataset tersebut akan mempengaruhi diagnosis kanker payudara perempuan. Suatu perempuan menderita diagnosis Maligna jika nilai variable-variabel prediktornya tinggi. Sedangkan, perempuan dengan diagnosis Benign (jinak) memiliki nilai variable-variabel predictor yang lebih rendah dibandingkan dengan diagnosis Maligna.

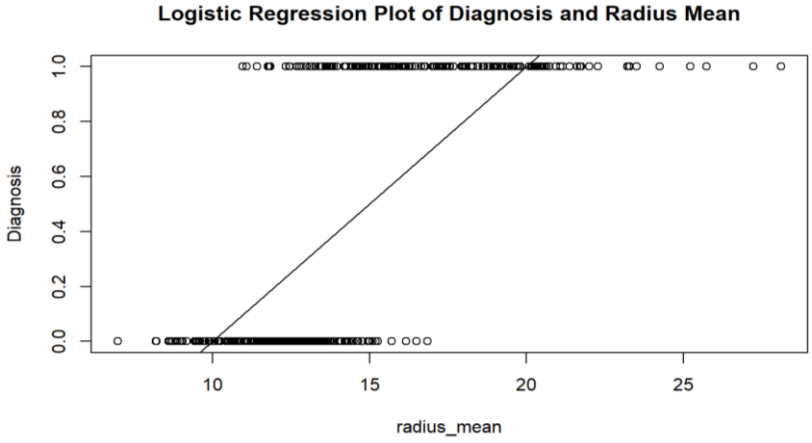
b. Construct the model regression you choose based on answer (a).

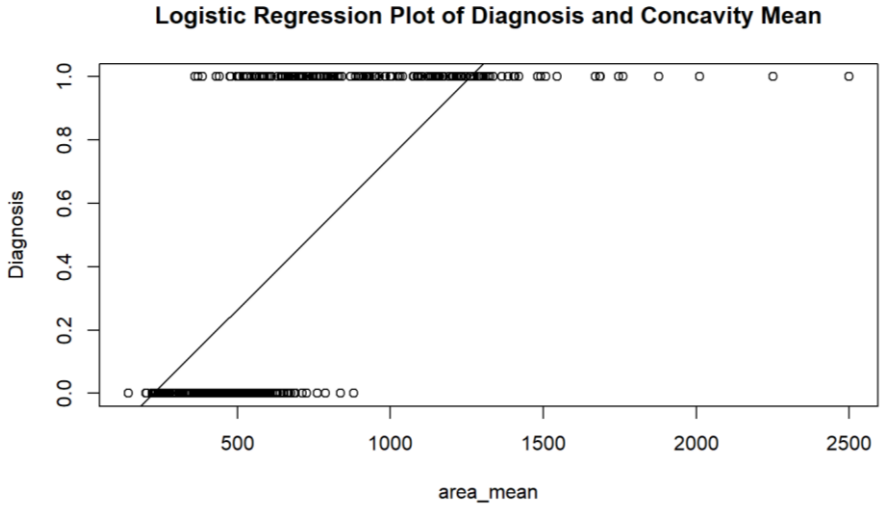
Jawab:

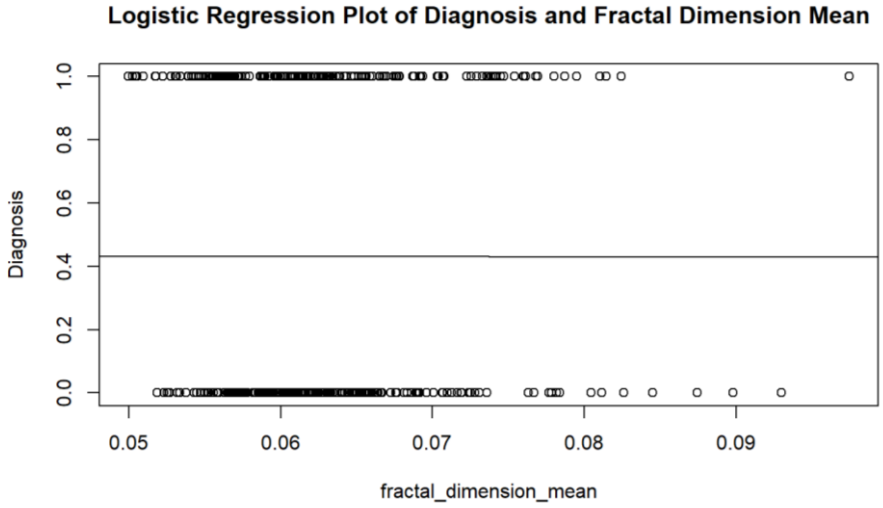
```
library(readxl)

df2 <- read_excel("Breast Cancer Wisconsin (Diagnostic).xlsx")
df2

summary(df2)

# Data-data diagnosis dijadikan numerik agar bisa di linear modelkan
Diagnosis <- ifelse(df2$diagnosis == "M", 1, 0)
Diagnosis <- as.numeric(Diagnosis)
Diagnosis

model2 <- lm(Diagnosis ~ ., data = df2)
model2

summary(model2)
```

Hasil plot

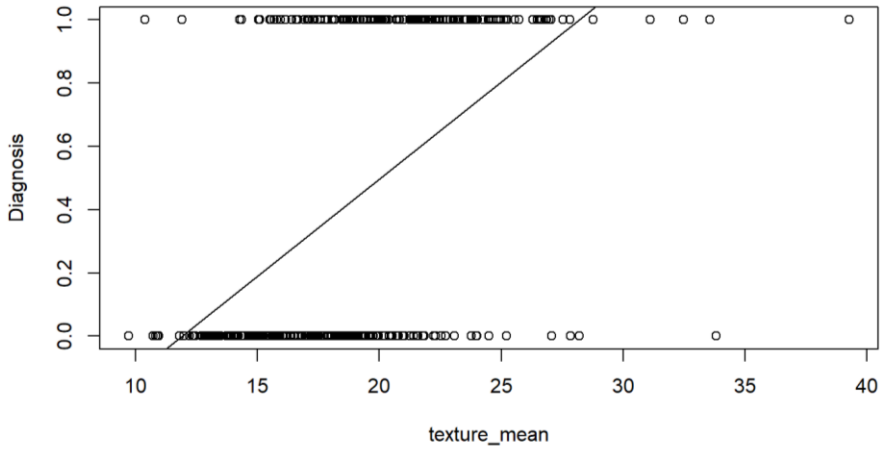



Logistic Regression Plot of Diagnosis and Radius Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
10
15
20
25
radius_mean

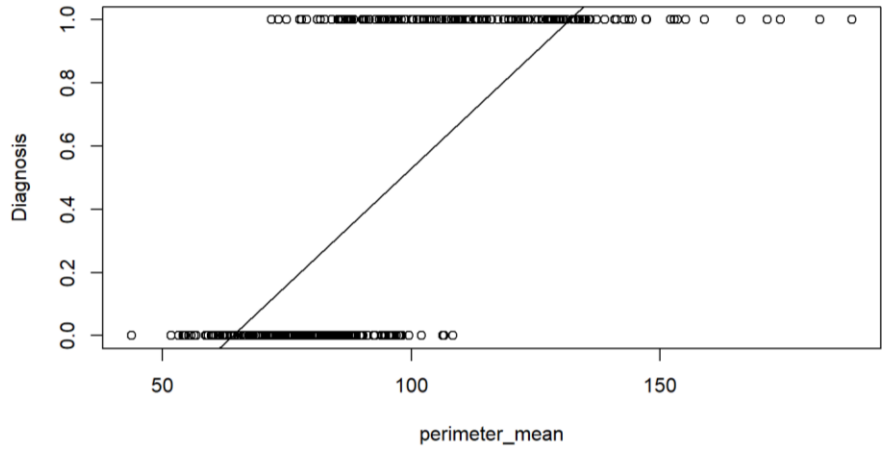

Logistic Regression Plot of Diagnosis and Concavity Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
500
1000
1500
2000
2500
area_mean


Logistic Regression Plot of Diagnosis and Fractal Dimension Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
0.05
0.06
0.07
0.08
0.09
fractal_dimension_mean

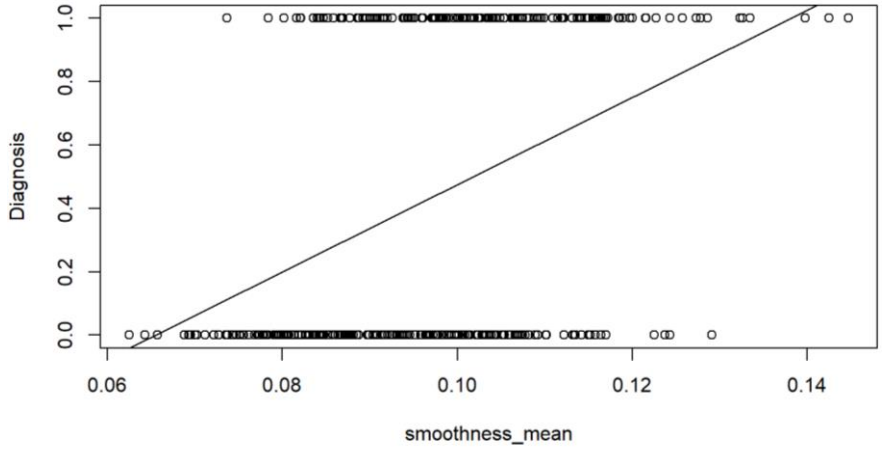
Logistic Regression Plot of Diagnosis and Texture Mean

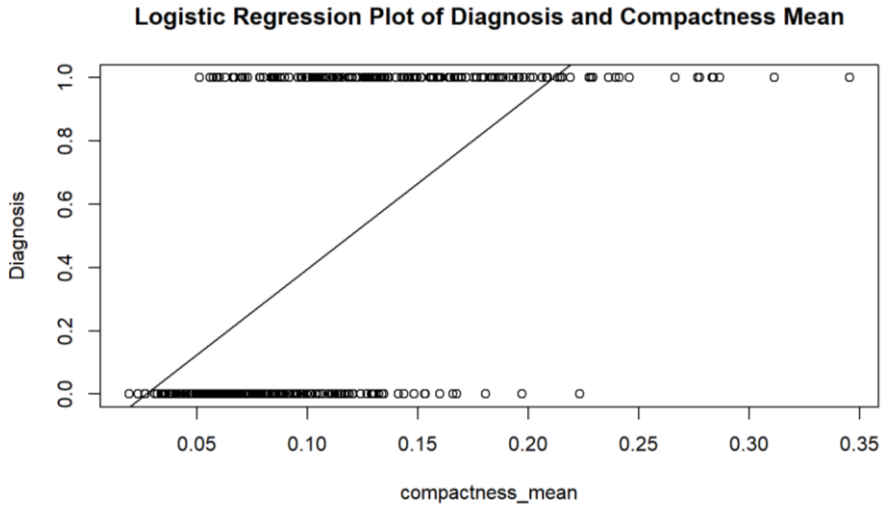


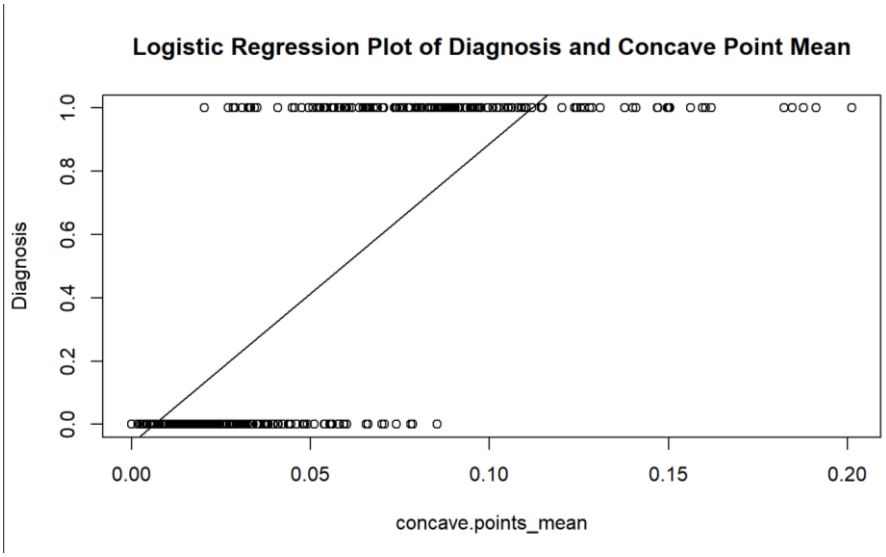
Logistic Regression Plot of Diagnosis and Perimeter Mean

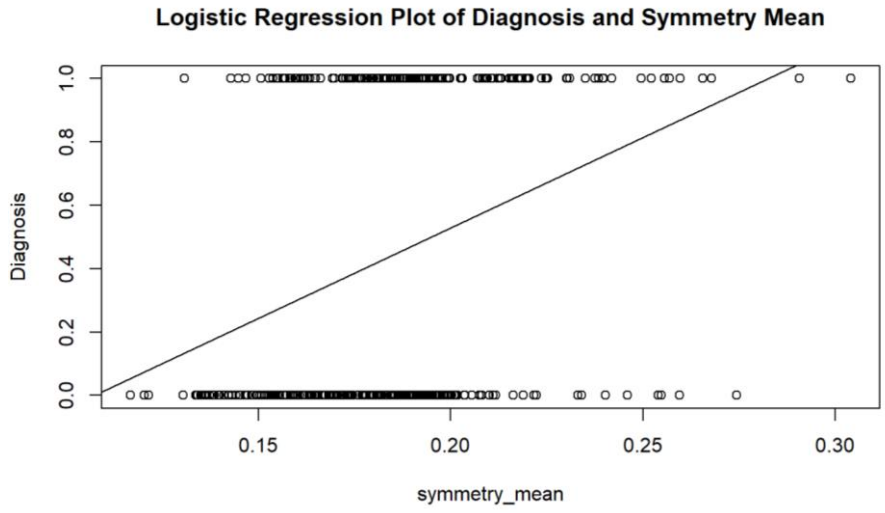


Logistic Regression Plot of Diagnosis and Smoothness Mean




Logistic Regression Plot of Diagnosis and Compactness Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
0.05
0.10
0.15
0.20
0.25
0.30
0.35
compactness_mean


Logistic Regression Plot of Diagnosis and Concave Point Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.05
0.10
0.15
0.20
concave.points_mean


Logistic Regression Plot of Diagnosis and Symmetry Mean
Diagnosis
0.0
0.2
0.4
0.6
0.8
1.0
0.15
0.20
0.25
0.30
symmetry_mean

c. Interpret your regression model. Give the details.

Jawab:

Hasil summary(model2)

Call:

lm(formula = Diagnosis ~ ., data = df2)

Residuals:

Min	1Q	Median	3Q	Max
-6.507e-16	-6.820e-17	-1.160e-17	3.200e-17	1.001e-14

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.797e-16	9.842e-16	-1.830e-01	0.8552
id	7.595e-26	2.397e-25	3.170e-01	0.7516
diagnosisM	1.000e+00	9.305e-17	1.075e+16	<2e-16
radius_mean	3.774e-16	2.871e-16	1.314e+00	0.1895
texture_mean	2.455e-18	7.596e-18	3.230e-01	0.7467
perimeter_mean	-6.238e-17	4.559e-17 -	-1.368e+00	0.1720
area_mean	1.057e-19	5.910e-19	1.790e-01	0.8582
smoothness_mean	-1.760e-15	3.200e-15 -	-5.500e-01	0.5826
compactness_mean	9.718e-16	2.239e-15	4.340e-01	0.6645
concavity_mean	-7.986e-16	1.058e-15	-7.540e-01	0.4510
concave.points_mean	6.629e-15	3.136e-15	2.113e+00	0.0352
symmetry_mean	2.297e-16	1.248e-15	1.840e-01	0.8541
fractal_dimension_mean	-3.123e-15	9.190e-15	-3.400e-01	0.7341

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

Residual standard error: 5.169e-16 on 387 degrees of freedom

Multiple R-squared: 1,

Adjusted R-squared: 1

F-statistic: 3.063e+31 on 12 and 387 DF,

p-value: < 2.2e-16

Dari hasil summary linear model untuk dataset tersebut dengan variable dependennya adalah diagnosis. Intercept standard error yang dihasilkan adalah 9.842e-16, estimate = -1.797e-16, t value = -1.830e-01, dan p-value(Pr(>|t|)) = 0.8552. Standard error terbesar dimiliki oleh variable diagnosis atau diagnosis Maligna yaitu sebesar 9.305e-17. Perkiraan terbesar dimiliki oleh compactness_mean sebesar 9.718e-16. t-value terbesar dimiliki oleh compactness_mean sebesar 4.340e-01. Serta, p-value(Pr(>|t|)) terbesar dimiliki oleh symmetry_mean sebesar 0.8541.

Selain itu, hasil summary(model2) juga menunjukkan bahwa linear model expenses memiliki residual standard error sebesar 5.169e-16 pada 387 derajat kebebasan, Multiple R-squared: 1, Adjusted R-squared: 1, F-statistic: 3.063e+31 pada 12 dan 387 DF, serta p-value lebih kecil dari 2.2e-16. Karena Multiple R-squared: 1 dan Adjusted R-squared: 1, maka model klasifikasi tersebut dibilang sangat baik.

Dari hasil plot diatas, bisa dijelaskan bahwa variable-variabel independen-nya berpengaruh besar kepada diagnosis kanker payudara seorang perempuan. Perempuan dengan diagnosis Maligna rata-rata memiliki nilai variable-variabel prediktor yang lebih besar daripada perempuan dengan kanker diagnosis Benign. Karena Maligna merupakan tumor yang ganas dan mengakibatkan kanker tinggi pada wanita. Sedangkan benign merupakan tumor yang jinak dan memiliki nilai-nilai variable prediktor yang lebih kecil dibandingkan dengan kebanyakan diagnosis Maligna.

d. From the dataset, use 70% of the data as training data and 30% of the data as a testing data then construct the classification model and define the accuracy of your classification model

Jawab:

```
dataf2 <- read.csv("Breast Cancer Wisconsin (Diagnostic).csv")
dataf2

library(caTools)

set.seed(1000)

# Menggunakan 70% data untuk training
```

```
split <- sample.split(dataf2, SplitRatio = 0.7)
dataf2_model <- dataf2[split,]
dataf2_test <- dataf2[!split,]

split
dataf2_model
dataf2_test

Diagnosis2 <- ifelse(dataf2_model$diagnosis == "M", 1, 0)
Diagnosis2 <- as.numeric(Diagnosis2)
Diagnosis2

model_logistic <- glm(Diagnosis2 ~ ., data = dataf2_model, family = 'binomial')
model_logistic

summary(model_logistic)

Diagnosis3 <- ifelse(dataf2_test$diagnosis == "M", 1, 0)
Diagnosis3 <- as.numeric(Diagnosis3)
Diagnosis3

prediction <- predict(model_logistic, dataf2_test, type="response")
# Cek probabilitas
prediction <- ifelse(prediction < 0.5, 0, 1)
prediction

# Construct classification model
model_pred <- prediction(prediction, Diagnosis3)
model_perform <- performance(model_pred, measure = "tpr", x.measure = "fpr")
auc <- performance(model_pred, measure = "auc")
auc <- auc@y.values[[1]]

plot(model_perform)
```

```
abline(a = 0, b = 1)
auc <- round(auc, 4)

# Menghitung akurasi
accuracy <- sum(Diagnosis3 == prediction)/length(Diagnosis3)
print(paste('Accuracy = ', accuracy))
```

Hasil akurasi = 1

Karena hasil akurasinya adalah 1 atau 100%, maka klasifikasi modelnya sangat akurat.