Tugas Akhir Praktikum Analisis Regresi 1

Abyoso Hapsoro Nurhadi

Jumat, 14 Juni 2019

Pra-Pengolahan

Pengantar

Notebook ini dibuat oleh Abyoso Hapsoro Nurhadi dengan NPM 1606884136 untuk memenuhi Tugas Akhir Praktikum Analisis Regresi 1. R yang digunakan adalah versi 3.5.3 dan RStudio yang digunakan adalah versi 1.1.463.

Library dan Setup

Pertama, setup beberapa hal untuk markdown.

```
# menangkap notebook
knitr::opts_chunk$set(cache=TRUE)

# set notasi non-saintifik untuk sesi R
options(scipen = 999)

# membersihkan environment
rm(list = ls())
```

Lalu load library yang diperlukan.

```
# jika package pacman tidak ada, install terlebih dahulu
if (!require("pacman")) install.packages("pacman")

# load library pacman
library(pacman)

# gunakan fungsi pacman untuk load library
# jika library tersebut tidak ada, install terlebih dahulu
p_load(mice, tidyverse, gridExtra, olsrr, gvlma)
```

Bersihkan console sebelum lanjut.

```
# membersihkan console
cat('\014')
```

Nomor 1

Diberikan kode sebagai berikut.

```
#loading data
data<-iris[,c(1,3)]
install.packages(mice)
#make missing data
library(mice)
alpha<-ampute(data, prop = 0.2, mech = "MAR",cont = TRUE,
bycases = TRUE, run = TRUE)
data1<-alpha$amp
names(data1)<-c("x","y")
#pada kasus ini, data1 memiliki 30 buah observasi yang mana
#salah satu dari x atau y missing.
#Salah satu teknik imputasi (mengisi missing data) adalah dengan menggunakan model regresi yang dibentu
#menghilangkan data missing
data2<-na.omit(data1)
#model y terhadap x
model1<-lm(data2$y~data2$x)
model1$coefficients
#y = -7.538727 + 1.931484*x yang ekuivalen dengan
#Model x terhadap y
model2<-lm(data2$x~data2$y)
model2$coefficients
#x = 4.3798285 + 0.3803727*y yang ekuivalen dengan
#y = 2.629*x - 11.51457
#dalam kasus ini, persamaan y terhadap x yang diberikan oleh kedua model berbeda
```

Apabila kode tersebut dijalankan, diperoleh hasil berikut.

```
## (Intercept)     data2$x
##   -7.674304    1.948682

## (Intercept)     data2$y
##   4.3694792   0.3884907
```

Jawaban Pertanyaan

1. Akan dijelaskan mengapa hal ini terjadi.

Asumsikan bahwa yang dimaksud dengan hal ini adalah kejadian berbedanya kedua model yang dihasilkan oleh model y terhadap x dan model x terhadap y. Misal model y terhadap x adalah model 1 dan model x terhadap y adalah model 2.

Dalam sebuah model regresi, variabel target diprediksi dengan variabel-variabel prediktornya. Artinya, model 1 memprediksi y dengan nilai x dan model 2 memprediksi x dengan nilai y. Hal inilah yang menyebabkan kedua model tidak sama, yakni variabel yang ingin diprediksi tidak sama. Walaupun persamaan model 2 diatur sedemikian sehingga menjadi persamaan untuk memprediksi y, secara konsep ini tidaklah benar karena persamaan ini seharusnya digunakan untuk memprediksi x bukan untuk memprediksi y. Sehingga persamaan model 1 dan model 2 berbeda.

2. Akan dijelaskan dan dilakukan model yang digunakan untuk mengimputasi data.

Perhatikan bahwa pada kode yang diberikan metode untuk membuat sebagian data missing adalah MAR yaitu Missing at Random. Data yang memiliki sifat missing value ini memiliki kecenderungan bahwa data yang hilang tidak berhubungan dengan data yang hilang, namun dengan beberapa data yang terobservasi. Dalam kasus ini, cukup aman untuk menghilangkan data dengan missing values karena tidak menciptakan bias yang signifikan. Namun kita tetap dapat mengimputasi data, dengan catatan melakukan imputasi belum tentu memberikan model yang lebih baik.

Dalam kasus soal ini, kode telah mengambil sebagian dari data iris, yaitu Sepal Length dan Petal Length yang dua-duanya bernilai real. Maka metode-metode imputasi yang biasa dilakukan adalah mean, median, dan modus. Namun kita juga dapat melakukan imputasi dengan metode Regresi Linier dan metode Multiple Imputation. Metode mean, median, dan modus akan menciptakan terlalu banyak data yang serupa dalam kasus regresi linier sederhana yang ada pada soal ini (karena hanya terdapat 1 fitur dan 1 target), sehingga ketiganya bukan merupakan pilihan yang baik. Secara teoritis, metode Regresi Linier memiliki hasil imputasi yang baik, namun cenderung untuk melakukan imputasi ini terlalu baik sehingga error standar deflasi serta diperlukan asumsi bahwa terdapat hubungan linier antara variabel fitur dengan target walaupun belum tentu ada. Artinya, bias yang diciptakan dalam mengimputasi dapat signifikan. Sehingga tersisa metode Multiple Imputation. Metode ini adalah metode yang paling baik untuk mengimputasi data kontinu karena bias yang diciptakan tidak signifikan.

Akan digunakan metode Multiple Imputation dengan model Markov Chain Monte Carlo. Metode ini sudah tersedia dalam package mice (Multivariate Imputation via Chained Equations) untuk digunakan. Metode MICE ini mengasumsikan bahwa missing data adalah MAR, yang sesuai dengan permasalahan ini.

```
# lihat pola missing value
md.pattern(data1)
```

```
##      y  x
## 120  1  1  0
## 16   1  0  1
## 14   0  1  1
##     14 16 30
```

Dari output tersebut, kita temukan bahwa sebanyak 121 observasi tidak memiliki missing value, 18 observasi memiliki missing value pada y, dan 11 observasi memiliki missing value pada x.

```
# imputasi regresi deterministik
imp <- mice(data1, method = "norm.predict", m = 5, seed = 123)

# bangun model prediktif dari kelima model
fit <- lm.mids(y ~ x, data = imp)
```

Bandingkan hasil model pada data tanpa imputasi (data2) dengan model pada data dengan imputasi missing values (data1_imp).

```
nonimp_model <- lm(y ~ x, data = data2)
summary(nonimp_model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2.42805 -0.59789 -0.07805  0.54504  2.62576
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  -7.6743     0.5859  -13.10 <0.0000000000000002 ***
## x             1.9487     0.1016   19.18 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8631 on 118 degrees of freedom
## Multiple R-squared:  0.757,  Adjusted R-squared:  0.755
## F-statistic: 367.7 on 1 and 118 DF,  p-value: < 0.00000000000000022
```

```
summary(pool(fit))
```

```
##             estimate  std.error statistic       df p.value
## (Intercept) -8.00497 0.46708280 -17.13822 146.0251       0
## x            2.01107 0.07915626  25.40633 146.0251       0
```

```
pool.r.squared(fit)
```

```
##           est     lo 95     hi 95 fmi
## R^2 0.8134803 0.7516681 0.8612744 NaN
```

```
pool.r.squared(fit, adjusted = TRUE)
```

```
##             est     lo 95     hi 95 fmi
## adj R^2 0.81222 0.7500557 0.8603097 NaN
```

Terlihat bahwa model dengan data yang diimputasi lebih baik dibandingkan dengan data yang tidak diimputasi pada kasus ini.

Nomor 2

Akan diaplikasikan analisis regresi linier dan diinterpretasikan setiap parameter pada model tersebut. Karena dibebaskan untuk mencari atau membuat dataset, penulis memutuskan untuk mencari dan menggunakan data eksternal. Data diambil dari https://github.com/kassambara/datarium/blob/master/data/marketing.rda yang dapat dipanggil melalui package datarium di dalam R.

Import Data

Baca data dan cek strukturnya.

```
# baca data
data("marketing", package = "datarium")

# cek 6 data pertama
head(marketing)
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

```
# cek eksistensi missing values
sapply(marketing, function(x) sum(is.na(x)))
```

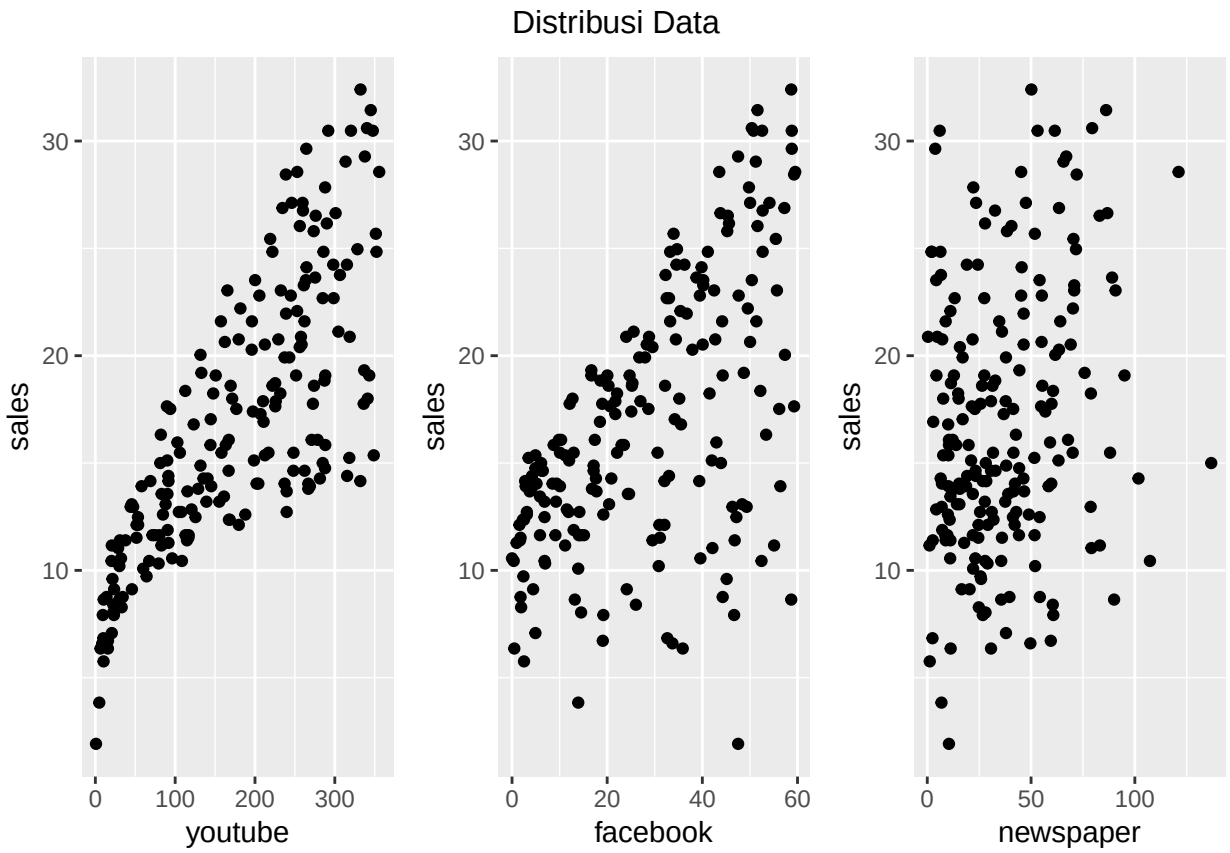
```
##   youtube  facebook newspaper     sales
##         0         0         0         0
```

Keterangan Data

Data marketing ini berisi 3 media periklanan (youtube, facebook, dan koran) dan sales yang dihasilkan. Entri sampel adalah budget periklanan dalam ribuan dollar bersama dengan sales. Terdapat 200 sampel dalam data.

Plot Persebaran Data

```
p1 <- ggplot(data = marketing, mapping = aes(x = youtube, y = sales)) + geom_point()
p2 <- ggplot(data = marketing, mapping = aes(x = facebook, y = sales)) + geom_point()
p3 <- ggplot(data = marketing, mapping = aes(x = newspaper, y = sales)) + geom_point()
grid.arrange(p1, p2, p3, nrow = 1, ncol = 3, top = "Distribusi Data")
```



Perhatikan bahwa ada pola kecenderungan naiknya sales bersama dengan naiknya budget periklanan youtube dan facebook. Tampak bahwa tidak ada pola kecenderungan dari newspaper terhadap sales. Hal ini dapat dianalisis lebih lanjut dengan melakukan pemilihan subset model terbaik.

Pemilihan Subset Model Terbaik

Misal:

- x1 adalah budget periklanan youtube
- x2 adalah budget periklanan facebook
- x3 adalah budget periklanan newspaper
- y adalah sales yang dihasilkan

```
names(marketing) <- c("x1", "x2", "x3", "y")
```

Akan ditentukan subset model terbaik untuk memprediksi sales.

```r
model <- lm(y ~ ., data = marketing)
ols_step_best_subset(model)
```

```
## Best Subsets Regression
## ----------------------
## Model Index    Predictors
## ----------------------
##      1         x1
##      2         x1 x2
##      3         x1 x2 x3
## ----------------------
##
##                                                 Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC         SBIC         SBC        MSEP        FPE       HSP       APC
## ---------------------------------------------------------------------------------------------------------------------------
##   1        0.6119      0.6099      0.6034    544.0814    1117.0200    545.4604    1126.9149    15.4456    15.4440    0.0776    0.3960
##   2        0.8972      0.8962      0.8925      2.0312     853.3227    285.8680     866.5160     4.1329     4.1319    0.0208    0.1059
##   3        0.8972      0.8956      0.8912      4.0000     855.2909    287.8779     871.7824     4.1747     4.1728    0.0210    0.1070
## ---------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Perhatikan bahwa performa model 3 menurun dari model 2 dilihat dari berbagai kriteria, seperti R-squared, adj R-squared, C(p), dan MSEP. Sehingga diperoleh cukup justifikasi untuk men-drop variabel newspaper.

```r
# drop kolom newspaper
marketing[, "x3"] <- list(NULL)

# cek ulang 6 data pertama
head(marketing)
```

```
##       x1    x2     y
## 1 276.12 45.36 26.52
## 2  53.40 47.16 12.48
## 3  20.64 55.08 11.16
## 4 181.80 49.56 22.20
## 5 216.96 12.96 15.48
## 6  10.44 58.68  8.64
```

Regresi Linier Tingkat Tinggi

Selanjutnya, akan dicoba beberapa subset-subset model order tinggi yang melibatkan faktor interaksi.

```
model <- lm(y ~ x1 + x2 + x1*x2 + I(x1^2) + I(x2^2), data = marketing)
ols_step_best_subset(model)
```

```
##          Best Subsets Regression
## ----------------------------------------
## Model Index    Predictors
## ----------------------------------------
##      1         x1:x2
##      2         x1 x1:x2
##      3         x1 I(x1^2) x1:x2
##      4         x1 x2 I(x1^2) x1:x2
##      5         x1 x2 I(x1^2) I(x2^2) x1:x2
## ----------------------------------------
##
##                                                    Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)        AIC           SBIC           SBC        MSEP      FPE       HSP       APC
## ---------------------------------------------------------------------------------------------------------------------------------------
##   1        0.9292      0.9288      0.9277    788.3200    776.8252      -73446.9732    786.7201    2.8189    2.8186    0.0142    0.0723
##   2        0.9661      0.9657      0.9648    277.5651    631.6464     -323550.5440    644.8397    1.3643    1.3639    0.0069    0.0350
##   3        0.9834      0.9832      0.9825     38.0643    490.1067    -1366709.4769    506.5983    0.6724    0.6721    0.0034    0.0172
##   4        0.9860      0.9857      0.9848      4.6223    458.6471    -1909986.3854    478.4370    0.5747    0.5743    0.0029    0.0147
##   5        0.9860      0.9857      0.9847      6.0000    460.0066    -1920336.5378    483.0948    0.5788    0.5782    0.0029    0.0148
## ---------------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
```

```
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Terlihat bahwa x2 tidak perlu dibawa ke tingkat tinggi karena menurunkan performa model. Sehingga untuk percobaan-percobaan selanjutnya akan difokuskan membawa x1 ke tingkat tinggi.

```
model <- lm(y ~ x1 + x2 + x1*x2 + I(x1^2) + I(x1^3) + I(x1^4) + I(x1^5) + I(x1^6), data = marketing)
ols_step_best_subset(model)
```

```
##                       Best Subsets Regression
## --------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------
##      1         x1:x2
##      2         x1 x1:x2
##      3         x1 I(x1^2) x1:x2
##      4         x1 I(x1^2) I(x1^3) x1:x2
##      5         x1 x2 I(x1^2) I(x1^3) x1:x2
##      6         x1 x2 I(x1^2) I(x1^3) I(x1^4) x1:x2
##      7         x1 x2 I(x1^2) I(x1^3) I(x1^4) I(x1^5) x1:x2
##      8         x1 x2 I(x1^2) I(x1^3) I(x1^4) I(x1^5) I(x1^6) x1:x2
## --------------------------------------------------------------------
##
##                                                  Subsets Regression Summary
## ------------------------------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC            SBIC           SBC        MSEP      FPE       HSP       APC
## ------------------------------------------------------------------------------------------------------------------------------------
##   1        0.9292      0.9288      0.9277    2523.2870    776.8252       -62882.1059    786.7201    2.8189    2.8186    0.0142    0.072
##   2        0.9661      0.9657      0.9648    1108.7481    631.6464       -277319.3552   644.8397    1.3643    1.3639    0.0069    0.035
##   3        0.9834      0.9832      0.9825     443.5767    490.1067       -1171958.1044  506.5983    0.6724    0.6721    0.0034    0.017
##   4        0.9874      0.9871      0.9863     294.6950    437.9041       -2016906.1842  457.6940    0.5181    0.5177    0.0026    0.013
##   5        0.9910      0.9907      0.9899     159.2994    373.2376       -3933299.2316  396.3259    0.3751    0.3747    0.0019    0.009
##   6        0.9931      0.9928       0.992      80.4734    322.2551       -6686381.3456  348.6416    0.2908    0.2904    0.0015    0.007
```

```
##   7        0.9941      0.9939      0.9928      43.3051    292.4588     -9191560.0381    322.1437    0.2506    0.2502    0.0013    0.006
##   8        0.9950      0.9948      0.9938       9.0000    259.6548    -13022420.7715    292.6380    0.2128    0.2123    0.0011    0.005
## ----------------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

```
model <- lm(y ~ x1 + x2 + x1*x2 + I(x1^2) + I(x1^2)*x2 + I(x1^3)*x2 + I(x1^4)*x2, data = marketing)
ols_step_best_subset(model)
```

```
##                             Best Subsets Regression
## --------------------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------------------
##      1         x1:x2
##      2         x1 x1:x2
##      3         x1 I(x1^2) x1:x2
##      4         x1 I(x1^2) I(x1^3) x1:x2
##      5         x1 x2 I(x1^2) I(x1^3) x1:x2
##      6         x1 x2 I(x1^2) I(x1^3) I(x1^4) x1:x2
##      7         x1 x2 I(x1^2) I(x1^3) I(x1^4) x1:x2 x2:I(x1^2)
##      8         x1 x2 I(x1^2) I(x1^3) I(x1^4) x1:x2 x2:I(x1^2) x2:I(x1^3)
##      9         x1 x2 I(x1^2) I(x1^3) I(x1^4) x1:x2 x2:I(x1^2) x2:I(x1^3) x2:I(x1^4)
## --------------------------------------------------------------------------------
##
##                                                 Subsets Regression Summary
## ----------------------------------------------------------------------------------------------------------------------------------------
##                       Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)         AIC          SBIC           SBC        MSEP      FPE       HSP       APC
## ----------------------------------------------------------------------------------------------------------------------------------------
##   1        0.9292      0.9288      0.9277    1777.6119    776.8252      -63569.3743    786.7201    2.8189    2.8186    0.0142    0.0723
##   2        0.9661      0.9657      0.9648     751.5122    631.6464      -280327.2846   644.8397    1.3643    1.3639    0.0069    0.0350
##   3        0.9834      0.9832      0.9825     269.2907    490.1067     -1184630.4289   506.5983    0.6724    0.6721    0.0034    0.0172
##   4        0.9874      0.9871      0.9863     161.7834    437.9041     -2038705.4677   457.6940    0.5181    0.5177    0.0026    0.0133
```

```
##   5        0.9910      0.9907      0.9899      64.0640    373.2376    -3975785.0981    396.3259    0.3751    0.3747    0.0019    0.0096
##   6        0.9931      0.9928       0.992       7.4018    322.2551    -6758577.4651    348.6416    0.2908    0.2904    0.0015    0.0074
##   7        0.9931      0.9929      0.9919       7.9137    322.7102    -6860338.8126    352.3951    0.2916    0.2910    0.0015    0.0075
##   8        0.9931      0.9928      0.9916       9.8794    324.6744    -6859087.3757    357.6576    0.2946    0.2939    0.0015    0.0075
##   9        0.9932      0.9929      0.9913      10.0000    324.7059    -6992105.6301    360.9873    0.2948    0.2940    0.0015    0.0075
## ----------------------------------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Terlihat dari kedua pemilihan subset di atas, x1 tingkat tinggi tanpa berinteraksi dengan x2 merupakan pilihan lebih baik karena C(p) lebih konsisten serta R-squared, adj R-squared, dan MSEP yang lebih bagus. Agar tidak terlalu membuat kompleks model dan tidak overfitting, akan digunakan model subset terbaik yaitu:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_1^3 + \beta_6 x_1^4 + \beta_7 x_1^5 + \beta_8 x_1^6$$


```r
model <- lm(y ~ x1 + x2 + x1*x2 + I(x1^2) + I(x1^3) + I(x1^4) + I(x1^5) + I(x1^6), data = marketing)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1 * x2 + I(x1^2) + I(x1^3) + I(x1^4) +
##     I(x1^5) + I(x1^6), data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81700 -0.24664  0.02012  0.24265  1.06693
##
## Coefficients:
##                         Estimate           Std. Error t value
## (Intercept)  2.43739707809046457  0.26402719961316629   9.232
## x1           0.28886262010729669  0.01861935098414323  15.514
## x2           0.04262588649596212  0.00359057859045419  11.872
## I(x1^2)     -0.00457781074069292  0.00044609209640314 -10.262
## I(x1^3)      0.00003874912009632  0.00000465098452449   8.331
```

```
## I(x1^4)     -0.00000017151539929  0.00000002363317105  -7.257
## I(x1^5)      0.00000000037668893  0.00000000005755745   6.545
## I(x1^6)     -0.00000000000032388  0.00000000000005375  -6.025
## x1:x2        0.00086637156781232  0.00001753160829073  49.418
##                            Pr(>|t|)
## (Intercept) < 0.0000000000000002 ***
## x1          < 0.0000000000000002 ***
## x2          < 0.0000000000000002 ***
## I(x1^2)     < 0.0000000000000002 ***
## I(x1^3)       0.0000000000000153 ***
## I(x1^4)       0.0000000000096555 ***
## I(x1^5)       0.0000000005350180 ***
## I(x1^6)       0.0000000085221637 ***
## x1:x2       < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4508 on 191 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9948
## F-statistic:  4775 on 8 and 191 DF,  p-value: < 0.00000000000000022
```


Dari hasil fitting tersebut, diperoleh model regresinya (setelah dibulatkan 4 angka di belakang desimal) sebagai berikut:

$$\hat{y} = 2.4374 + 0.2889x_1 + 0.0426x_2 + 0.0009x_1x_2 - 0.0046x_1^2$$

Ternyata walaupun terlihat pada step pemilihan subset model terbaik ada kenaikan R-squared serta penurunan MSEP, pengaruh koefisien regresi tingkat tinggi dari x1 (youtube) tidak lebih signifikan dari 0.0001.

Karena ini dan untuk menghindari terlalu mengkomplikasi model, akan digunakan model yang menjelaskan persamaan di atas, yaitu:

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + \beta_4x_1^2$$

```r
model <- lm(y ~ x1 + x2 + x1*x2 + I(x1^2), data = marketing)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1 * x2 + I(x1^2), data = marketing)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9939 -0.3563 -0.0080  0.4557  1.4023
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  6.164474685  0.231200457  26.663 < 0.0000000000000002 ***
## x1           0.050920326  0.002232391  22.810 < 0.0000000000000002 ***
## x2           0.035162337  0.005900624   5.959         0.0000000117 ***
## I(x1^2)     -0.000091454  0.000005745 -15.920 < 0.0000000000000002 ***
## x1:x2        0.000897173  0.000028884  31.061 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7485 on 195 degrees of freedom
## Multiple R-squared:  0.986,  Adjusted R-squared:  0.9857
## F-statistic:  3432 on 4 and 195 DF,  p-value: < 0.00000000000000022
```

Dari hasil fitting tersebut, diperoleh model regresinya (setelah dibulatkan 4 angka di belakang desimal) sebagai berikut:

$$\hat{y} = 6.1645 + 0.0509x_1 + 0.0352x_2 + 0.0009x_1x_2 - 0.0001x_1^2$$

Model ini memiliki adj R-squared 0.9857 yang sudah jauh lebih memuaskan dibandingkan model subset terbaik awal yang tidak menginkorporasikan tingkat tinggi maupun interaksi dengan adj R-squared 0.8962. Semua variabel yang digunakan berguna dengan kepercayaan 99% karena Pr(>F) < 0.01 untuk setiap variabel.

Interpetasi

Setiap x2 bertambah 1 (budget periklanan facebook bertambah 1000 dollar), secara rata-rata y bertambah 0.1409x1 - 0.0001x1^2 (sales bertambah nilai tersebut dikali 1000 dollar).

Karena ada suku x1 kuadrat, tidak dapat diberikan interpretasi untuk penambahan x1.

Uji Asumsi

Terdapat 10 asumsi-asumsi dalam regresi linier, namun secara umum terdapat 5 asumsi yang terpenting yang dapat dipanggil dengan package gvlma (Global Validation of Linear Models Assumptions).

```
gvlma(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x1 * x2 + I(x1^2), data = marketing)
##
## Coefficients:
## (Intercept)           x1           x2      I(x1^2)        x1:x2
##  6.16447469   0.05092033   0.03516234  -0.00009145   0.00089717
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model)
##
##                      Value  p-value                   Decision
## Global Stat        4042.859 0.000000 Assumptions NOT satisfied!
## Skewness            297.700 0.000000 Assumptions NOT satisfied!
## Kurtosis           3733.467 0.000000 Assumptions NOT satisfied!
## Link Function        10.474 0.001211 Assumptions NOT satisfied!
## Heteroscedasticity    1.218 0.269718    Assumptions acceptable.
```

Dari output-output tersebut, ternyata hanya 1 dari 5 asumsi yang terpenuhi. Secara detil:

1. Penolakan Global Stat mengindikasikan adanya hubungan non-linier antara satu atau lebih dari prediktor dengan target. Walaupun ini ditolak, dari penjabaran yang sudah dilakukan dianggap sudah cukup terjustifikasi untuk hanya mengambil suku hingga tingkat 2 dari x1. Mungkin penolakan ini terjadi mengindikasikan perlunya penggunaan tingkat yang lebih tinggi dari x1.
2. Penolakan Skewness mengindikasikan bahwa data sebaiknya ditransformasi. Namun karena saya tidak ahli dengan R, saya belum mengetahui bagaimana caranya.
3. Penolakan Kurtosis mengindikasikan hal yang sama dengan penolakan Skewness.
4. Penolakan Link Function mengindikasikan bahwa perlu digunakan bentuk alternatif dari Generalized Linear Model seperti Regresi Binomial. Namun karena mata kuliah ini adalah mata kuliah Regresi Linier, hal ini tidak akan diusahakan untuk dicapai.
5. Penerimaan Heterokedastisitas mengindikasikan bahwa variansi residual dari model konstan dari seluruh range nilai prediktor-prediktor.

Dari argumen tersebut, anggap 1, 4, dan 5 sudah memenuhi. Untuk pengembangan model ini, dapat diusahakan untuk memenuhi asumsi-asumsi kedua dan ketiga dengan melakukan transformasi data. Serta perlu dikaji ulang apakah suku tingkat tinggi diperlukan pada data yang sudah ditransformasi, serta hingga seberapa besar tingkat baik yang terkait.

Namun karena adj R-squared sudah memuaskan pada kasus ini, maka sejauh tugas akhir dan mata kuliah ini, saya cukupkan hingga di sini.

Post-Pengolahan

Penutup

Sekian pekerjaan saya, semoga sudah memenuhi tugas yang telah diberikan. Terima kasih.