ITERA

Modul 2 Praktikum
Statistika Sains Data

Regresi Linier Sederhana dan Regresi Linier Berganda

Program Studi Sains Data
Fakultas Sains
Institut Teknologi Sumatera

2024

A. Tujuan Praktikum

1. Mahasiswa mampu menaksir model regresi linier sederhana dan regresi linier berganda menggunakan software RStudio.
2. Mahasiswa mampu menguji signifikansi parameter dari persamaan regresi linier sederhana dan regresi linier berganda yang telah diolah dengan RStudio.
3. Mahasiswa mampu menentukan kualitas dari model regresi yang terbentuk.

B. Teori Dasar

I. Regresi linier

merupakan model regresi yang menggunakan garis lurus untuk menggambarkan hubungan antar variabel . Ia menemukan garis yang paling sesuai dengan data Anda dengan mencari nilai koefisien regresi yang meminimalkan kesalahan total model. Ada dua jenis utama regresi linier:
Regresi linier sederhana hanya menggunakan satu variabel independent
Regresi linier berganda menggunakan dua atau lebih variabel independent.

II. Contoh model regresi linier sederhana dan regresi linier berganda dalam pemrograman R

Mulailah dengan mengunduh R dan RStudio . Kemudian buka RStudio dan klik **File > File Baru > R Script** . Saat kita menjalani setiap langkah , Anda dapat menyalin dan menempelkan kode dari kotak teks langsung ke skrip Anda. Untuk menjalankan kode, **sorot baris yang ingin Anda jalankan** dan klik tombol **Run** di kanan atas editor teks (atau tekan **ctrl + enter** pada keyboard). Untuk menginstal paket yang Anda perlukan untuk analisis, jalankan kode ini (Anda hanya perlu melakukan ini sekali):

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("broom")
install.packages("ggpubr")
```

Selanjutnya, muat paket ke lingkungan R Anda dengan menjalankan kode ini (Anda perlu melakukan ini setiap kali memulai ulang R):

```
library(ggplot2)
library(dplyr)
library(broom)
library(ggpubr)
```

Langkah 1: Muat data ke R

Data contoh ini dapat anda **download** disni. Ikuti empat langkah berikut untuk setiap kumpulan data:

a. Di RStudio, buka File > Impor kumpulan data > Dari Teks (basis) .
b. Pilih file data yang telah Anda unduh (income.data atau heart.data), dan jendela **Impor Kumpulan Data** akan muncul.
c. Di jendela **Data Frame** , Anda akan melihat kolom **X** (indeks) dan kolom yang mencantumkan data untuk masing-masing variabel (pendapatan dan kebahagiaan atau bersepeda , merokok , dan penyakit jantung).
d. Klik tombol **Impor** dan file akan muncul di tab **Lingkungan** Anda di sisi kanan atas layar RStudio. Setelah Anda memuat data, periksa apakah data telah dibaca dengan benar menggunakan summary ().

Regresi sederhana

```
summary(income.data)
```

Karena kedua variabel kita bersifat kuantitatif , saat kita menjalankan fungsi ini, kita akan melihat tabel di konsol dengan ringkasan data numerik. Ini memberi tahu kita nilai minimum, **median , mean ,** dan maksimum dari variabel independen (pendapatan) dan variabel dependen (kebahagiaan):

```
       X                income          happiness    
 Min.   :  1.0   Min.   :1.506   Min.   :0.266  
 1st Qu.:125.2   1st Qu.:3.006   1st Qu.:2.266  
 Median :249.5   Median :4.424   Median :3.473  
 Mean   :249.5   Mean   :4.467   Mean   :3.393  
 3rd Qu.:373.8   3rd Qu.:5.992   3rd Qu.:4.503  
 Max.   :498.0   Max.   :7.482   Max.   :6.863  
```

Regresi berganda

```
summary(heart.data)
```

Sekali lagi, karena variabelnya bersifat kuantitatif, menjalankan kode akan menghasilkan ringkasan numerik dari data untuk variabel independen (merokok dan bersepeda) dan variabel dependen (penyakit jantung):

```
       X                biking          smoking      heart.disease    
 Min.   :  1.0   Min.   : 1.119   Min.   : 0.5259   Min.   : 0.5519  
 1st Qu.:125.2   1st Qu.:20.205   1st Qu.: 8.2798   1st Qu.: 6.5137  
 Median :249.5   Median :35.824   Median :15.8146   Median :10.3853  
 Mean   :249.5   Mean   :37.788   Mean   :15.4350   Mean   :10.1745  
 3rd Qu.:373.8   3rd Qu.:57.853   3rd Qu.:22.5689   3rd Qu.:13.7240  
 Max.   :498.0   Max.   :74.907   Max.   :29.9467   Max.   :20.4535  
```

Langkah 2: Pastikan data Anda memenuhi asumsi

Kita dapat menggunakan R untuk memeriksa apakah data kita memenuhi empat asumsi utama regresi linier .

Regresi sederhana

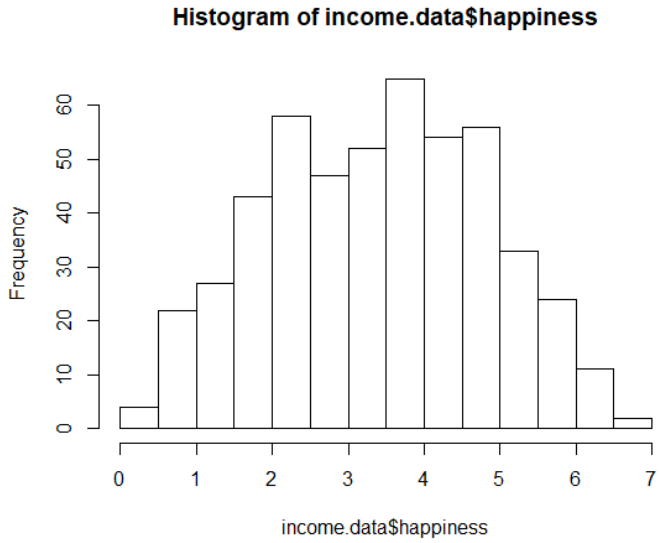
1. **Independensi observasi** (alias tidak ada autokorelasi)

   Karena kita hanya mempunyai satu variabel bebas dan satu variabel terikat, kita tidak perlu menguji adanya hubungan tersembunyi antar variabel. Jika Anda mengetahui bahwa Anda mempunyai autokorelasi dalam variabel (yaitu beberapa observasi pada subjek uji yang sama), maka jangan lanjutkan dengan regresi linier sederhana! Gunakan model terstruktur, seperti model efek campuran linier.

2. **Normalitas**

   Untuk memeriksa apakah variabel terikat mengikuti distribusi normal , gunakan hist()fungsi.

   ```
   hist(income.data$happiness)
   ```

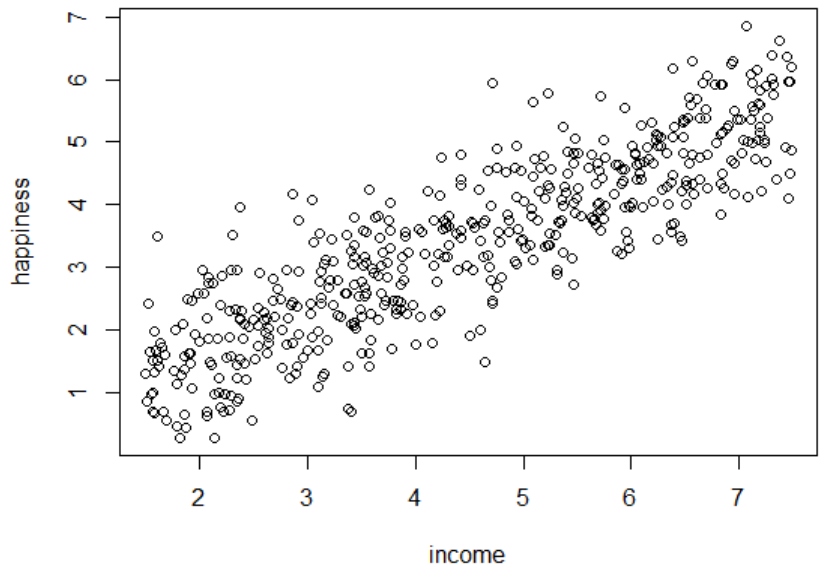


Pengamatannya secara kasar berbentuk lonceng (lebih banyak observasi di tengah distribusi, lebih sedikit di bagian ekor), sehingga kita dapat melanjutkan dengan regresi linier.

3. **Linearitas**

Hubungan antara variabel independen dan dependen harus linier. Kita dapat mengujinya secara visual dengan plot sebar untuk melihat apakah sebaran titik data dapat digambarkan dengan garis lurus.

```
plot(happiness ~ income, data = income.data)
```



Hubungannya terlihat linier, sehingga kita dapat melanjutkan dengan model linier.

4. **Homoskedastisitas** (alias homogenitas varians)

   Artinya kesalahan prediksi tidak berubah secara signifikan sepanjang rentang prediksi model. Kita dapat menguji asumsi ini nanti, setelah memasang model linier

Regresi Berganda

1. **Independensi observasi** (alias tidak ada autokorelasi)

   Gunakan cor() fungsi tersebut untuk menguji hubungan antara variabel independen dan pastikan variabel tersebut tidak berkorelasi terlalu tinggi.

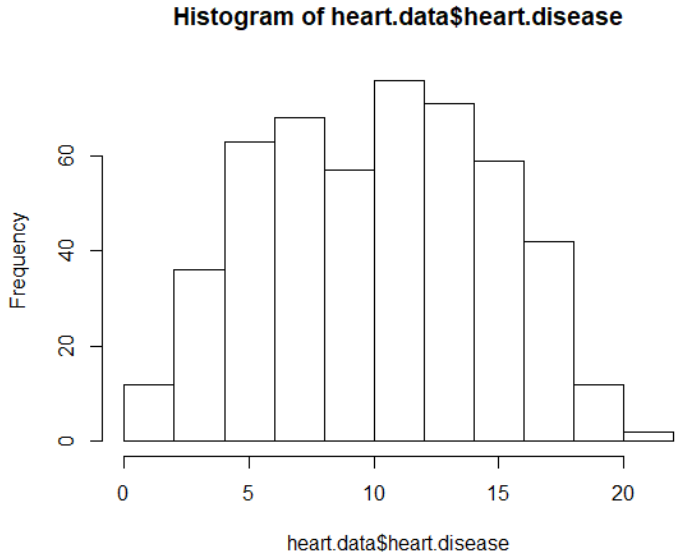
```
cor(heart.data$biking, heart.data$smoking)
```

   Saat kita menjalankan kode ini, outputnya adalah 0,015. Korelasi antara bersepeda dan merokok kecil (0,015 hanya merupakan korelasi 1,5%), jadi kami dapat memasukkan kedua parameter tersebut ke dalam model kami.

2. **Normalitas**

   Gunakan hist()fungsi tersebut untuk menguji apakah variabel terikat Anda mengikuti distribusi normal .

```
hist(heart.data$heart.disease)
```

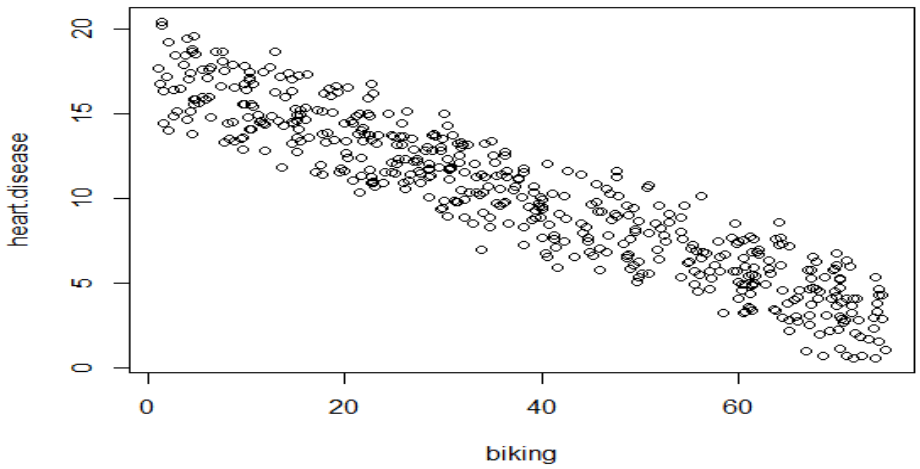


   Distribusi observasi kira-kira berbentuk lonceng, sehingga kita dapat melanjutkan dengan regresi linier.

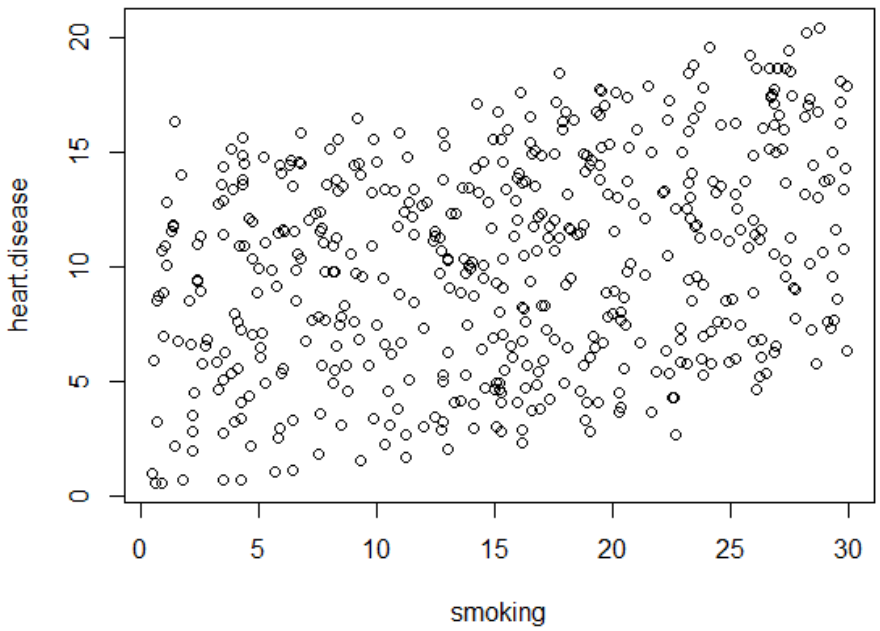
3. **Linearitas**

Kita dapat memeriksanya menggunakan dua diagram sebar: satu untuk bersepeda dan penyakit jantung, dan satu lagi untuk merokok dan penyakit jantung.

```
plot(heart.disease ~ biking, data=heart.data)
```




```
plot(heart.disease ~ smoking, data=heart.data)
```



Meskipun hubungan antara merokok dan penyakit jantung kurang jelas, namun hubungan tersebut masih tampak linier. Kita bisa melanjutkan dengan regresi linier.

4. **Homoskedastisitas**
   Kami akan memeriksanya setelah kami membuat modelnya.

   Langkah 3: Lakukan analisis regresi linier
   Sekarang setelah Anda menentukan bahwa data Anda memenuhi asumsi, Anda dapat melakukan analisis regresi linier untuk mengevaluasi hubungan antara variabel independen dan dependen.

   Regresi sederhana: pendapatan dan kebahagiaan
   Mari kita lihat apakah ada hubungan linier antara pendapatan dan kebahagiaan dalam survei kami terhadap 500 orang dengan pendapatan berkisar antara $15k hingga $75k, di mana kebahagiaan diukur pada skala 1 hingga 10. Untuk melakukan analisis regresi linier sederhana dan memeriksa hasilnya, Anda perlu menjalankan dua baris kode. Baris kode pertama membuat model linier, dan baris kedua mencetak ringkasan model:

```
income.happiness.lm <- lm(happiness ~ income, data = income.data)

summary(income.happiness.lm)
```

   Outputnya terlihat seperti ini:

```
Call:
lm(formula = happiness ~ income, data = income.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02479 -0.48526  0.04078  0.45898  2.37805

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.20427    0.08884   2.299   0.0219 *
income       0.71383    0.01854  38.505   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7181 on 496 degrees of freedom
Multiple R-squared:  0.7493,    Adjusted R-squared:  0.7488
F-statistic:  1483 on 1 and 496 DF,  p-value: < 2.2e-16
```

   Tabel keluaran ini pertama-tama menyajikan persamaan model, kemudian merangkum residu model (lihat langkah 4).
   Bagian **Koefisien** menunjukkan:

1. Estimasi (**Estimasi**) untuk parameter model – nilai titik potong y (dalam hal ini 0,204) dan estimasi pengaruh pendapatan terhadap kebahagiaan (0,713).
2. Kesalahan standar dari nilai estimasi (**Std. Error**).
3. Statistik uji (nilai **t ,** dalam hal ini *t* statistik).
4. Nilai *p* (**Pr(>| t |)**) , alias probabilitas menemukan statistik *t* yang diberikan jika hipotesis nol tidak ada hubungan benar.

   Tiga baris terakhir adalah diagnostik model – hal terpenting yang perlu diperhatikan adalah **nilai *p*** (inilah 2,2e-16, atau hampir nol), yang akan menunjukkan apakah model tersebut cocok dengan data. Dari hasil tersebut dapat dikatakan bahwa terdapat **hubungan positif yang signifikan** antara pendapatan dan kebahagiaan (*p* value <0,001), dengan peningkatan kebahagiaan sebesar 0,713 unit (+/- 0,01) untuk setiap peningkatan pendapatan satu unit.

Regresi berganda: bersepeda, merokok, dan penyakit jantung

Mari kita lihat apakah ada hubungan linier antara bersepeda ke tempat kerja, merokok, dan penyakit jantung dalam survei imajiner kami terhadap 500 kota. Tingkat bersepeda ke tempat kerja berkisar antara 1 dan 75%, tingkat merokok antara 0,5 dan 30%, dan tingkat penyakit jantung antara 0,5% dan 20,5%. Untuk menguji hubungan tersebut, pertama-tama kami memasang model linier dengan penyakit jantung sebagai variabel terikat dan bersepeda serta merokok sebagai variabel bebas. Jalankan dua baris kode ini:

```
heart.disease.lm<-lm(heart.disease ~ biking + smoking, data = heart.data)

summary(heart.disease.lm)
```

Outputnya terlihat seperti ini:

```
Call:
lm(formula = heart.disease ~ biking + smoking, data = heart.data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1789 -0.4463  0.0362  0.4422  1.9331

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.984658   0.080137  186.99   <2e-16 ***
biking      -0.200133   0.001366 -146.53   <2e-16 ***
smoking      0.178334   0.003539   50.39   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.654 on 495 degrees of freedom
Multiple R-squared:  0.9796,    Adjusted R-squared:  0.9795
F-statistic: 1.19e+04 on 2 and 495 DF,  p-value: < 2.2e-16
```

Estimasi dampak bersepeda terhadap penyakit jantung adalah -0,2, sedangkan estimasi dampak merokok adalah 0,178. Artinya, setiap peningkatan 1% orang yang bersepeda ke tempat kerja, terdapat korelasi penurunan kejadian penyakit jantung sebesar 0,2%. Sedangkan setiap peningkatan 1% jumlah perokok menyebabkan peningkatan angka penyakit jantung sebesar 0,178%. Kesalahan standar untuk koefisien regresi ini sangat kecil, dan statistik *t* sangat besar (masing-masing -147 dan 50,4). Nilai *p* mencerminkan kesalahan kecil dan statistik *t* besar .

Langkah 4: Periksa homoskedastisitas

Sebelum melanjutkan dengan visualisasi data, kita harus memastikan bahwa model kita sesuai dengan asumsi homoskedastisitas model linier.

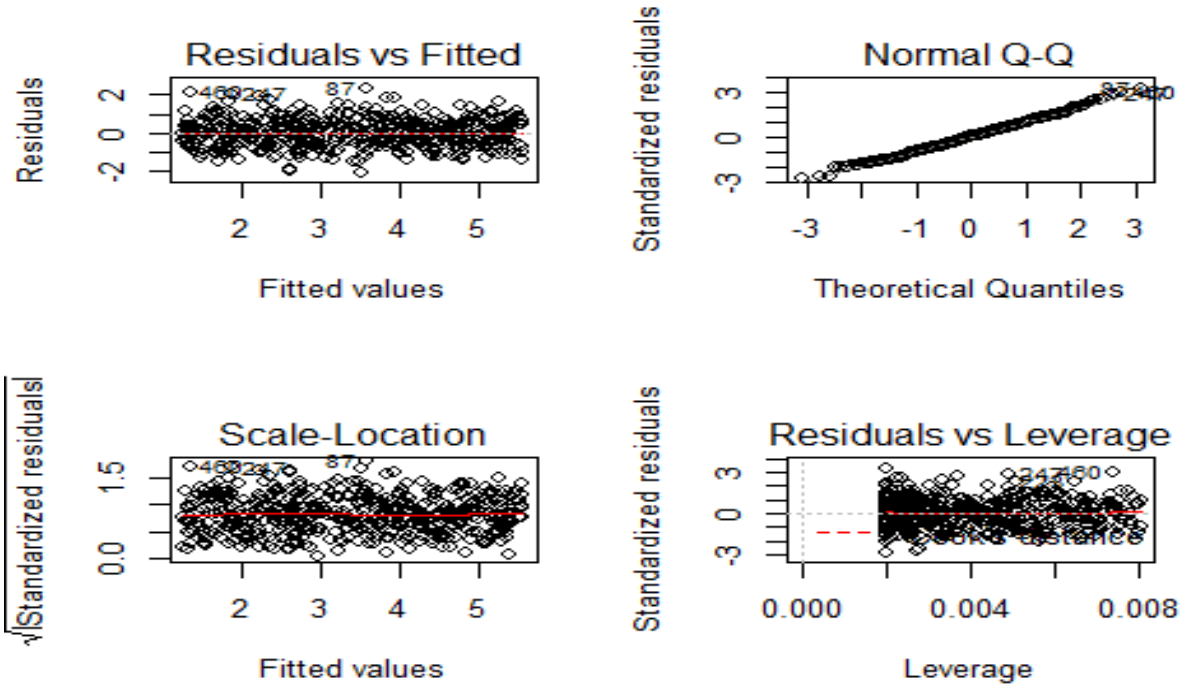
Regresi sederhana

Kita dapat menjalankannya plot(income.happiness.lm)untuk memeriksa apakah data yang diamati memenuhi asumsi model kita:

```
par(mfrow=c(2,2))
plot(income.happiness.lm)
par(mfrow=c(1,1))
```

Perhatikan bahwa par(mfrow())perintah tersebut akan membagi jendela **Plots** menjadi jumlah baris dan kolom yang ditentukan dalam tanda kurung. Jadi par(mfrow=c(2,2))bagilah menjadi dua baris

dan dua kolom. Untuk kembali memplot satu grafik di seluruh jendela, atur kembali parameternya dan ganti (2,2) dengan (1,1). Ini adalah plot yang dihasilkan oleh kode:



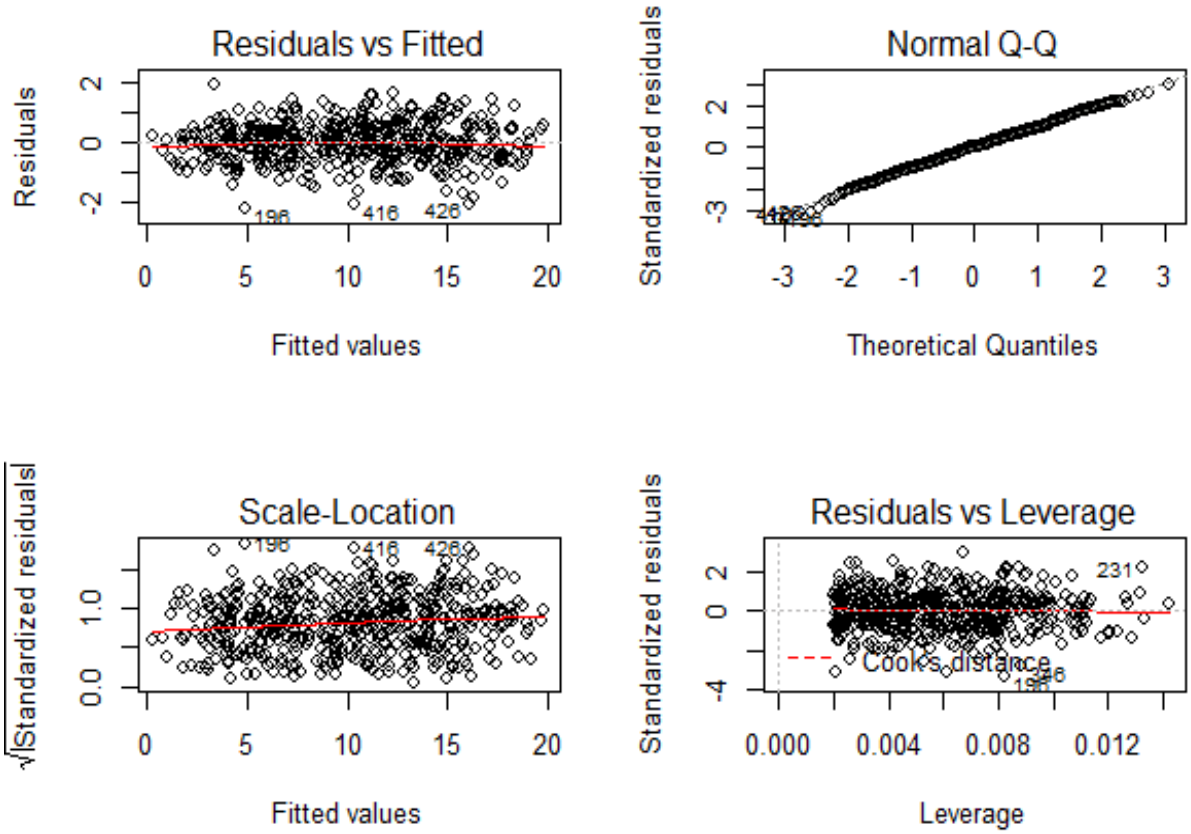
Residual adalah varians yang tidak dapat dijelaskan . Nilai tersebut tidak persis sama dengan kesalahan model, namun dihitung berdasarkan kesalahan tersebut, jadi melihat bias pada residu juga akan menunjukkan adanya bias pada kesalahan tersebut. Hal terpenting yang harus diperhatikan adalah garis merah yang mewakili rata-rata residu pada dasarnya horizontal dan berpusat di sekitar nol. Artinya tidak ada outlier atau bias pada data yang membuat regresi linier menjadi tidak valid. Pada **plot Q-Q Normal** di kanan atas, kita dapat melihat bahwa residu nyata dari model kita membentuk garis satu-satu yang hampir sempurna dengan residu teoretis dari model sempurna. Berdasarkan sisa-sisa tersebut, kita dapat mengatakan bahwa model kita memenuhi asumsi homoskedastisitas.

Regresi berganda

Kita harus memeriksa apakah model kita benar-benar sesuai dengan datanya, dan tidak ada variasi besar dalam kesalahan model, dengan menjalankan kode ini:

```
par(mfrow=c(2,2))
plot(heart.disease.lm)
par(mfrow=c(1,1))
```

Outputnya terlihat seperti ini:



Seperti halnya regresi sederhana kita, residunya tidak menunjukkan bias, sehingga kita dapat mengatakan bahwa model kita sesuai dengan asumsi homoskedastisitas.

Langkah 5: Visualisasikan hasilnya dengan grafik

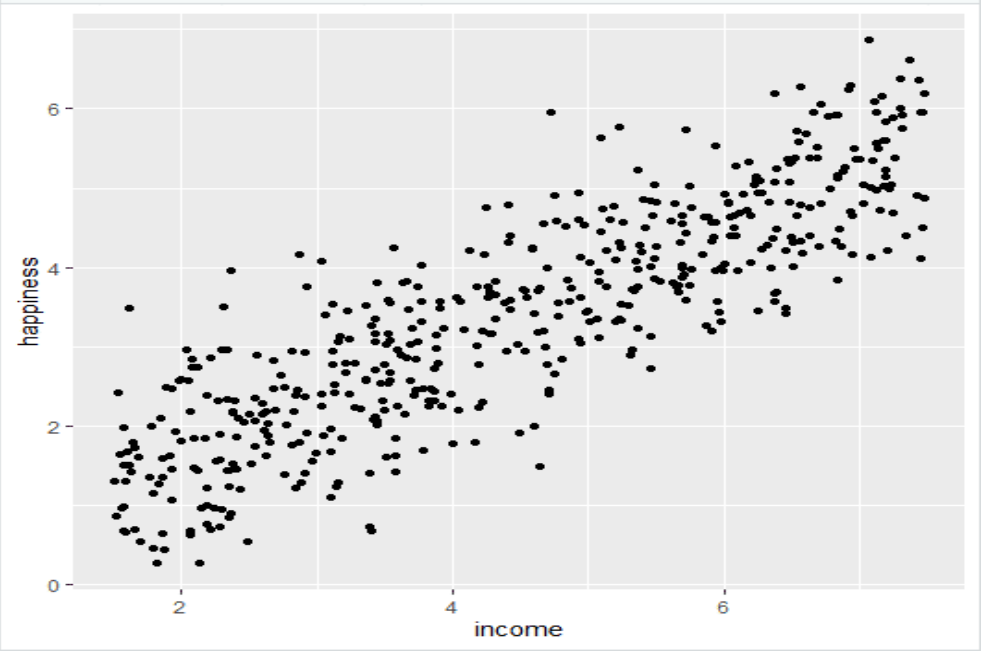
Selanjutnya, kita dapat memplot data dan garis regresi dari model regresi linier kita sehingga hasilnya dapat dibagikan.

Regresi sederhana

Ikuti 4 langkah untuk memvisualisasikan hasil regresi linier sederhana Anda.

1. **Plot titik data pada grafik**

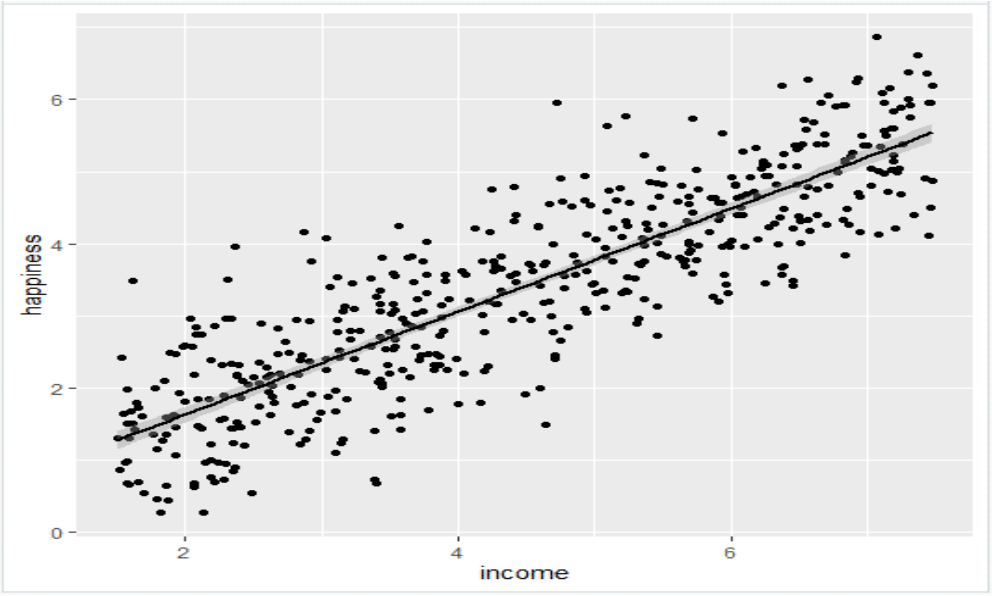
```
income.graph<-ggplot(income.data, aes(x=income, y=happiness))+
                     geom_point()
income.graph
```



2. Tambahkan garis regresi linier ke data yang diplot

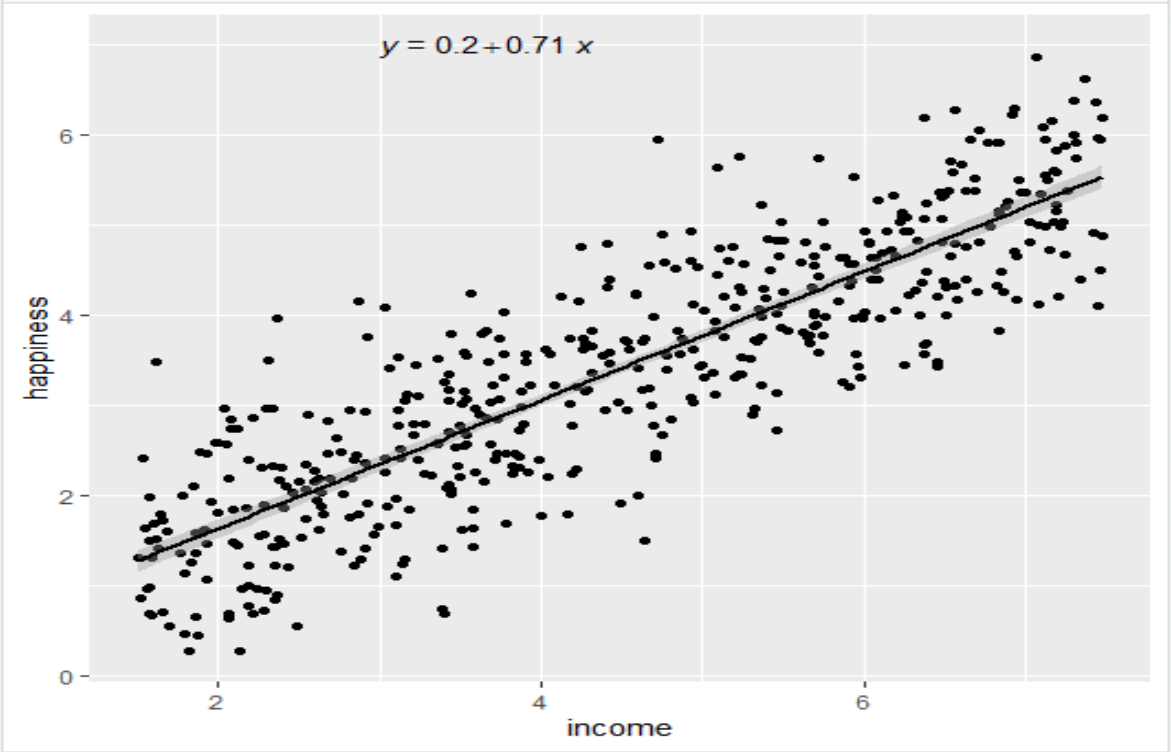
Tambahkan garis regresi menggunakan geom_smooth()dan mengetik lmsebagai metode Anda untuk membuat garis. Ini akan menambahkan garis regresi linier serta kesalahan standar estimasi (dalam hal ini +/- 0,01) sebagai garis abu-abu terang yang mengelilingi garis:

```
income.graph <- income.graph + geom_smooth(method="lm", col="black")

income.graph
```



3. **Tambahkan persamaan untuk garis regresi.**

```
income.graph <- income.graph +
  stat_regline_equation(label.x = 3, label.y = 7)

income.graph
```

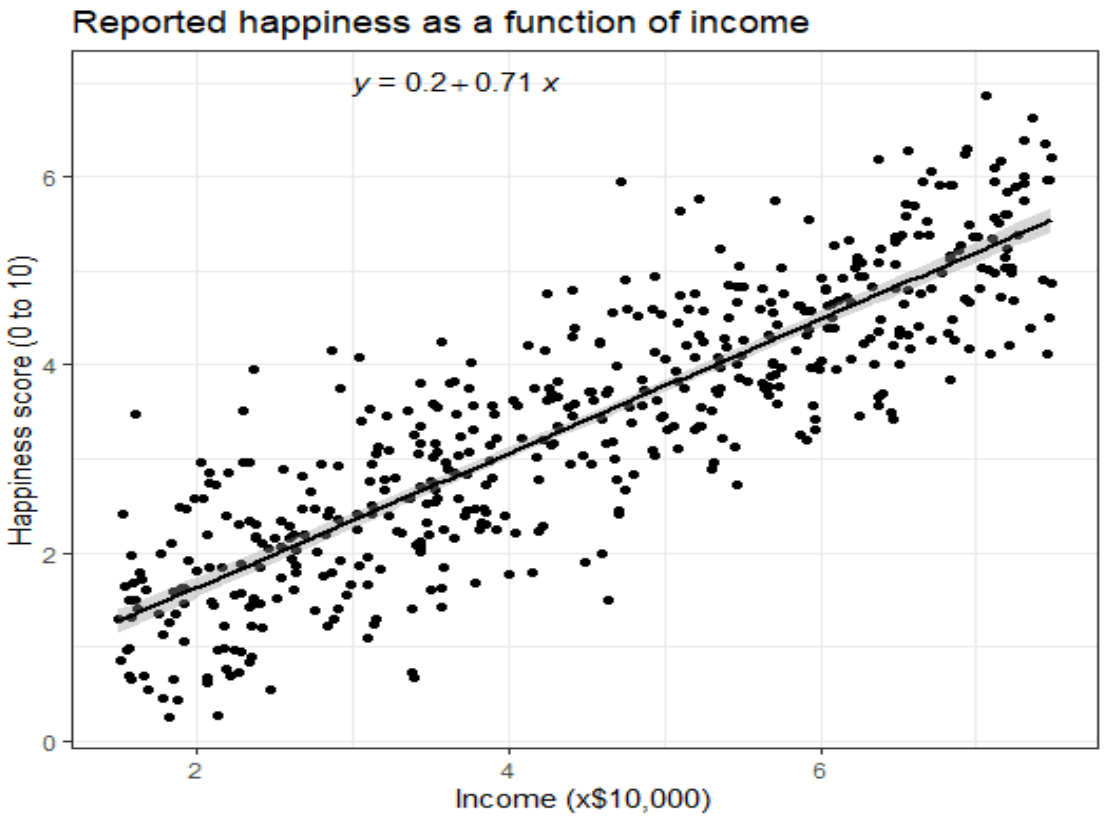


4. **Siapkan grafik untuk dipublikasikan**

Kita dapat menambahkan beberapa parameter gaya menggunakan theme_bw()dan membuat label khusus menggunakan labs().

```
income.graph +
  theme_bw() +
  labs(title = "Reported happiness as a function of income",
       x = "Income (x$10,000)",
       y = "Happiness score (0 to 10)")
```

Ini menghasilkan grafik selesai yang dapat Anda sertakan dalam makalah Anda:



Regresi berganda

Langkah visualisasi regresi berganda lebih sulit dibandingkan regresi sederhana, karena sekarang kita memiliki dua prediktor. Kami akan mencoba metode yang berbeda: menggambarkan hubungan antara bersepeda dan penyakit jantung pada tingkat perokok yang berbeda. Dalam contoh ini, merokok akan diperlakukan sebagai faktor dengan tiga tingkatan, hanya untuk tujuan menampilkan hubungan dalam data kita. Ada 7 langkah yang harus diikuti.

1. **Buat kerangka data baru dengan informasi yang diperlukan untuk memplot model**
   Gunakan fungsi expand.grid()untuk membuat kerangka data dengan parameter yang Anda berikan. Dalam fungsi ini kita akan:

- Buat urutan dari nilai terendah hingga tertinggi dari data bersepeda yang Anda amati;
- Pilih nilai minimum, rata-rata, dan maksimum dari merokok, untuk membuat 3 tingkat merokok yang dapat digunakan untuk memprediksi tingkat penyakit jantung.

```
plotting.data<-expand.grid(
  biking = seq(min(heart.data$biking), max(heart.data$biking), length.out=30),
    smoking=c(min(heart.data$smoking), mean(heart.data$smoking),
max(heart.data$smoking)))
```

2. **Prediksikan nilai penyakit jantung berdasarkan model linier Anda**

   Selanjutnya kita akan menyimpan nilai 'prediksi y' kita sebagai kolom baru di kumpulan data yang baru saja kita buat.

```
plotting.data$predicted.y <- predict.lm(heart.disease.lm, newdata=plotting.data)
```

3. **Bulatkan angka merokok menjadi dua decimal**

   Ini akan membuat legenda lebih mudah dibaca nantinya.

```
plotting.data$smoking <- round(plotting.data$smoking, digits = 2)
```

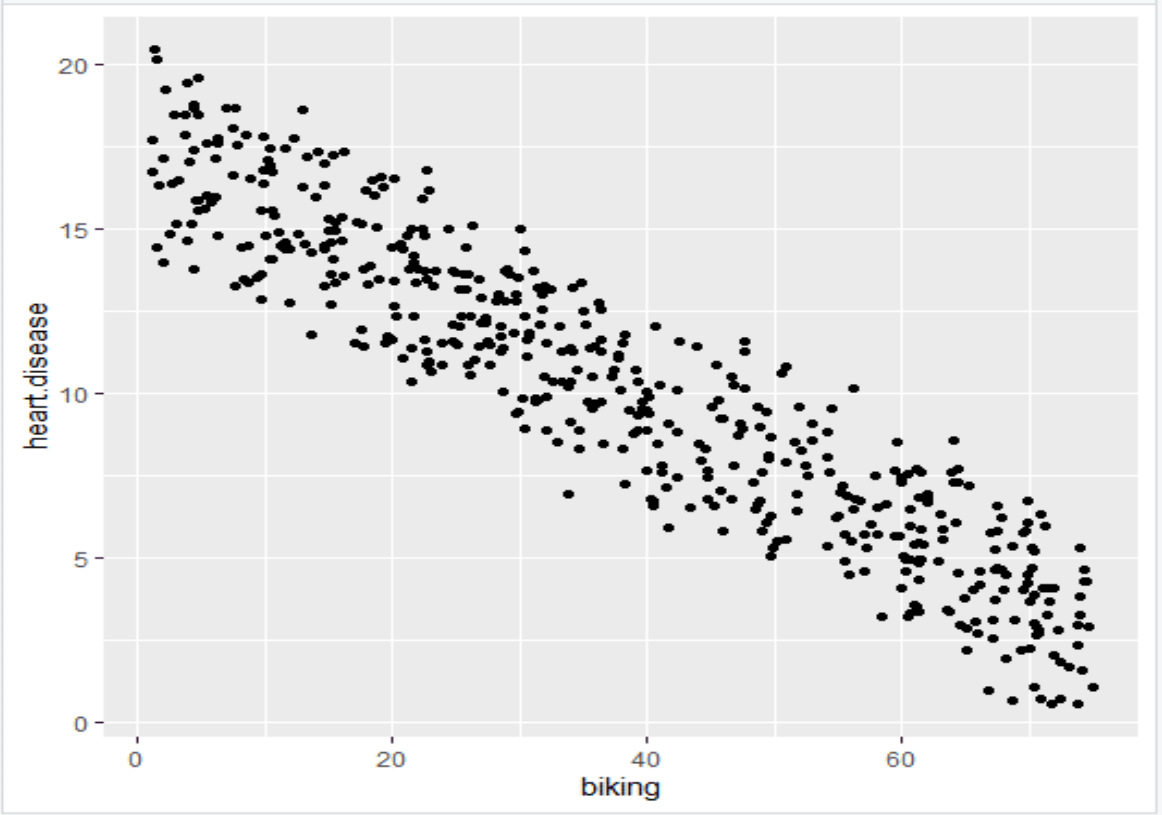
4. **Ubah variabel 'merokok' menjadi factor**

   Hal ini memungkinkan kami untuk menggambarkan interaksi antara bersepeda dan penyakit jantung pada masing-masing dari tiga tingkat kebiasaan merokok yang kami pilih.

```
plotting.data$smoking <- as.factor(plotting.data$smoking)
```

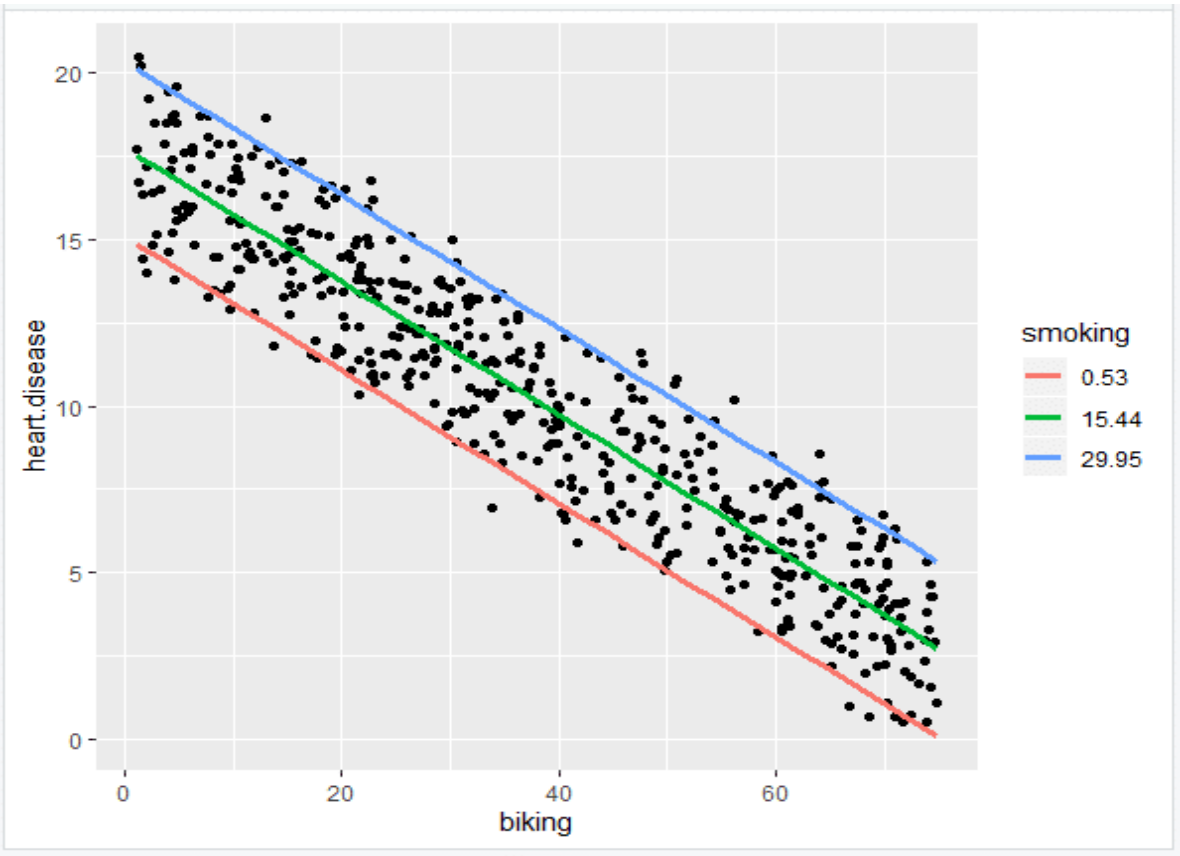
5. **Plot data asli**

```
heart.plot <- ggplot(heart.data, aes(x=biking, y=heart.disease)) +
  geom_point()

heart.plot
```



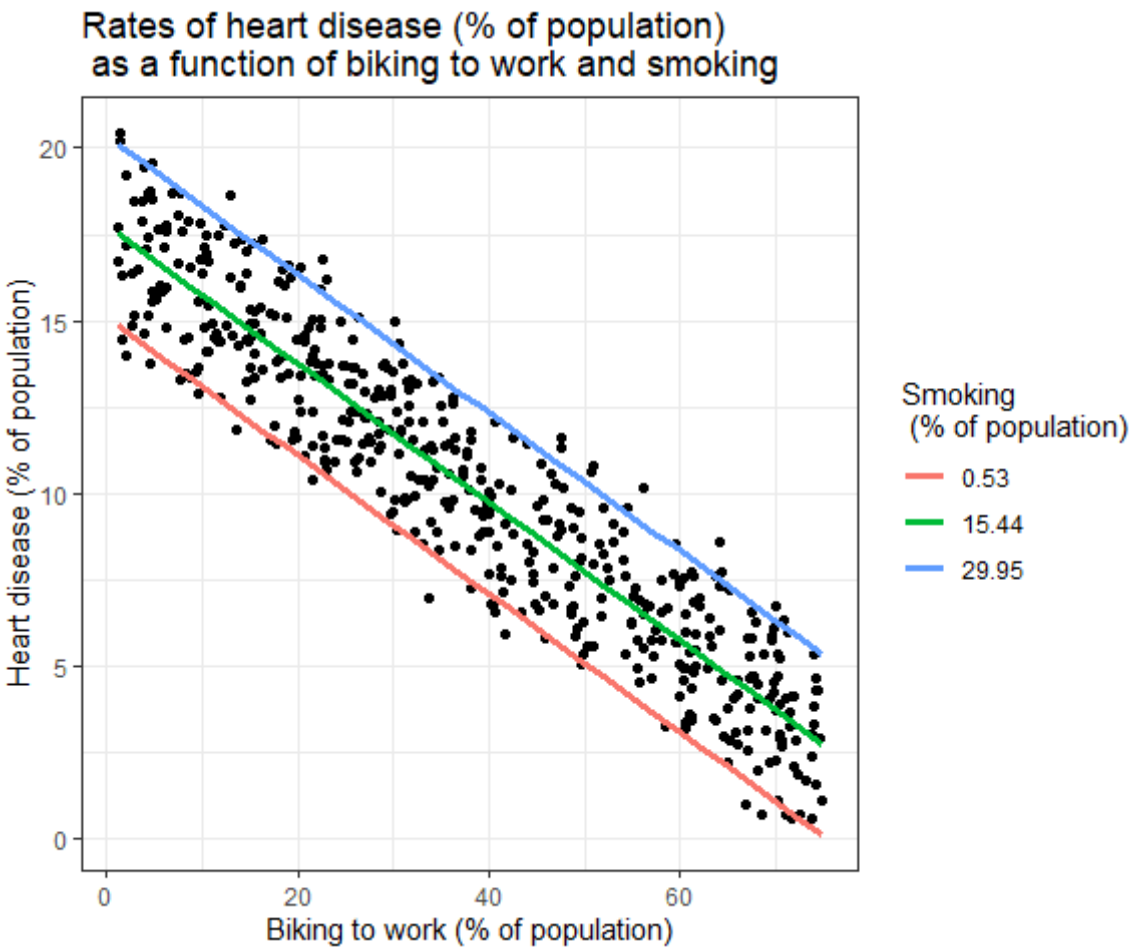
6. Tambahkan garis regresi

```
heart.plot <- heart.plot +
  geom_line(data=plotting.data, aes(x=biking, y=predicted.y, color=smoking), size=1.25)

heart.plot
```



7. Siapkan grafik untuk dipublikasikan

```
heart.plot <-
heart.plot +
  theme_bw() +
  labs(title = "Rates of heart disease (% of population) \n as a function of biking to
work and smoking",
       x = "Biking to work (% of population)",
       y = "Heart disease (% of population)",
       color = "Smoking \n (% of population)")

heart.plot
```



Karena grafik ini memiliki dua koefisien regresi, fungsinya stat_regline_equation()tidak akan berfungsi di sini. Namun jika kita ingin menambahkan model regresi ke grafik, kita dapat melakukannya seperti ini:

```
heart.plot + annotate(geom="text", x=30, y=1.75, label=" = 15 + (-0.2*biking) +
(0.178*smoking)")
```

Langkah 6: Laporkan hasil Anda

Melaporkan hasil regresi linier sederhana

Kami menemukan hubungan yang signifikan antara pendapatan dan kebahagiaan (p <0,001, R^2 = 0,73 ± 0,0193), dengan peningkatan kebahagiaan yang dilaporkan sebesar 0,73 unit untuk setiap peningkatan pendapatan sebesar $10.000.

Melaporkan hasil regresi linier berganda

Dalam survei kami terhadap 500 kota, kami menemukan hubungan yang signifikan antara frekuensi bersepeda ke tempat kerja dan frekuensi penyakit jantung, serta frekuensi merokok dan frekuensi penyakit jantung (masing-masing p <0 dan p <0,001).
Secara khusus kami menemukan penurunan 0,2% (± 0,0014) frekuensi penyakit jantung untuk setiap peningkatan 1% bersepeda, dan peningkatan 0,178% (± 0,0035) frekuensi penyakit jantung untuk setiap peningkatan 1% kebiasaan merokok.

C. Latihan Praktikum

Dalam hal ini, akan dilakukan praktikum berupa model regresi linier sederhana dan regresi linier berganda sebagai berikut.

Data ini bisa diperoleh di link berikut ini

Download Data

Memanggil Package

```
library(lmtest)
```

Import Data

```
rumah <- read.table("C:/Data/data rumah.txt", header=TRUE)
View(rumah)
str(rumah)
```

Output yang dihasilkan berupa

```
## 'data.frame':    522 obs. of  13 variables:
##  $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sales_price: int  360000 340000 250000 205500 275500 248000 229900 150000 195000 160000 ...
##  $ X1         : int  3032 2058 1780 1638 2196 1966 2216 1597 1622 1976 ...
##  $ X2         : int  4 4 4 4 4 4 3 2 3 3 ...
##  $ X3         : int  4 2 3 2 3 3 2 1 2 3 ...
##  $ X4         : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ X5         : int  2 2 2 2 2 5 2 1 2 1 ...
##  $ X6         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ X7         : int  1972 1976 1980 1963 1968 1972 1972 1955 1975 1918 ...
##  $ X8         : int  2 2 2 2 2 2 2 2 3 3 ...
##  $ X9         : int  1 1 1 1 7 1 7 1 1 1 ...
##  $ X10        : int  22221 22912 21345 17342 21786 18902 18639 22112 14321 32358 ...
##  $ X11        : int  0 0 0 0 0 0 0 0 0 0 ...
```

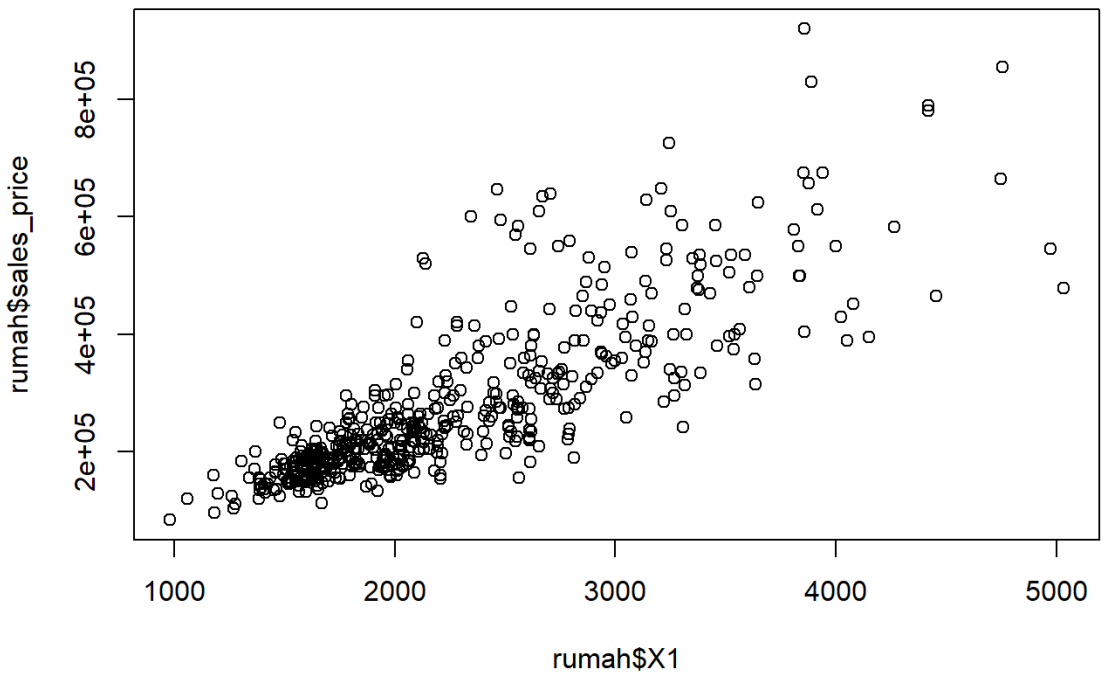
```
head(rumah)
```

Output yang dihasilkan

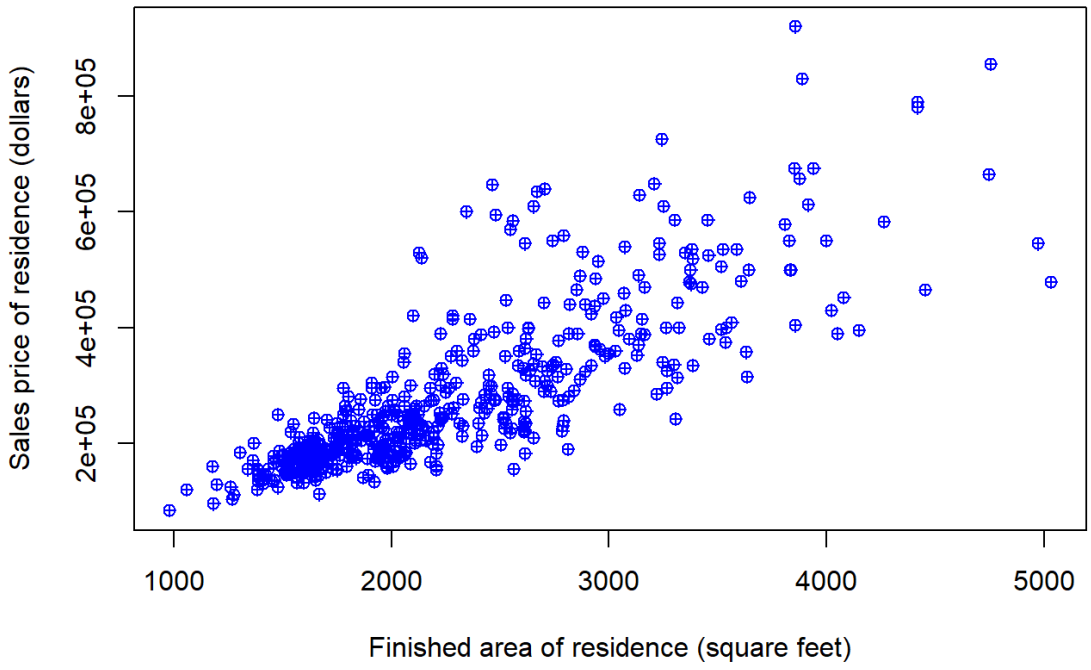
```
##   ID sales_price   X1 X2 X3 X4 X5 X6   X7 X8 X9   X10 X11
## 1  1      360000 3032  4  4  1  2  0 1972  2  1 22221   0
## 2  2      340000 2058  4  2  1  2  0 1976  2  1 22912   0
## 3  3      250000 1780  4  3  1  2  0 1980  2  1 21345   0
## 4  4      205500 1638  4  2  1  2  0 1963  2  1 17342   0
## 5  5      275500 2196  4  3  1  2  0 1968  2  7 21786   0
## 6  6      248000 1966  4  3  1  5  1 1972  2  1 18902   0
```

Visualisasi Data

```
#scatter plot
plot(rumah$X1, rumah$sales_price)
```




```
plot(rumah$X1, rumah$sales_price,
     xlab="Finished area of residence (square feet)",
     ylab="Sales price of residence (dollars)",
     col="blue",pch=10)
```



Menghitung Korelasi

```
cor(rumah$X1, rumah$sales_price)
```

Output yang dihasilkan

```
## [1] 0.8194701
```

Membuat Model Regresi Linier Sederhana

```
#model regresi linear sederhana dengan
#dependent var: sales_price
#independent var: X1
lm(sales_price ~ 1 + X1, data=rumah)
```

Output yang dihasilkan

```
##
## Call:
## lm(formula = sales_price ~ 1 + X1, data = rumah)
##
## Coefficients:
## (Intercept)           X1
##      -81433          159
```

```
lm(sales_price ~ X1, data=rumah)
```

```
##
## Call:
## lm(formula = sales_price ~ X1, data = rumah)
##
## Coefficients:
## (Intercept)           X1
##      -81433          159
```

```
model1 <- lm(sales_price ~ 1 + X1, data=rumah)
summary(model1)
```

Output yang dihasilkan

```
##
## Call:
## lm(formula = sales_price ~ 1 + X1, data = rumah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -239405  -39840   -7641   23515  388362
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -81432.946  11551.846  -7.049 5.74e-12 ***
## X1             158.950      4.875  32.605  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79120 on 520 degrees of freedom
## Multiple R-squared:  0.6715, Adjusted R-squared:  0.6709
## F-statistic:  1063 on 1 and 520 DF,  p-value: < 2.2e-16
```

```
anova(model1)
```

Output yang dihasilkan

```
## Analysis of Variance Table
##
## Response: sales_price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## X1          1 6.6555e+12 6.6555e+12  1063.1 < 2.2e-16 ***
## Residuals 520 3.2554e+12 6.2604e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Pendugaan Parameter

```
##pendugaan parameter
#dependent var: sales_price
#independent var: X1
y <- rumah$sales_price
X <- cbind(1,rumah$X1)
#rumus betaduga = (X'X)^-1*X'y
betaduga <- solve(t(X)%*%X)%*%t(X)%*%y
betaduga
```

Output yang dihasilkan

```
##             [,1]
## [1,] -81432.9464
## [2,]    158.9502
```

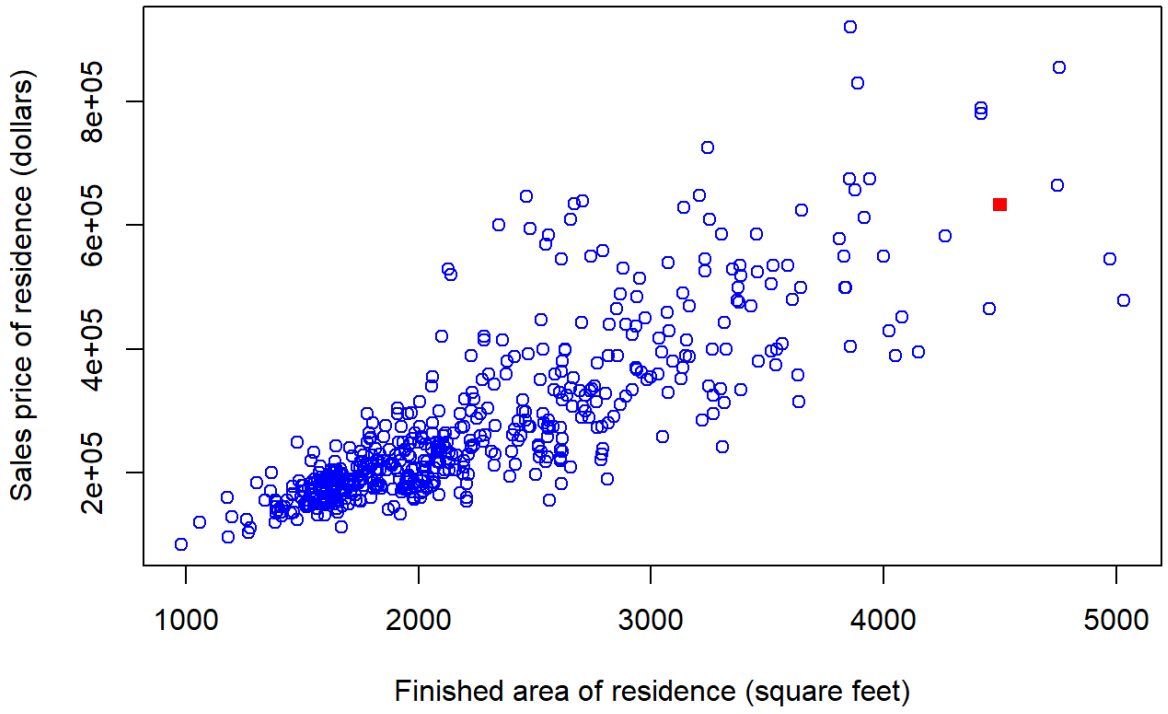
Memprediksi Harga Rumah

```
#memprediksi harga rumah seluas 4500
rumahku <- c(4500)
rumahku <- data.frame(rumahku)
colnames(rumahku) <- c("X1")
predict(model1, newdata=rumahku)
```

Output yang dihasilkan

```
##        1
## 633843.1
```

```
plot(rumah$X1, rumah$sales_price,
     xlab="Finished area of residence (square feet)",
     ylab="Sales price of residence (dollars)",
     col="blue")
points(rumahku$X1, predict(model1, newdata=rumahku), col="red", pch=15)
```



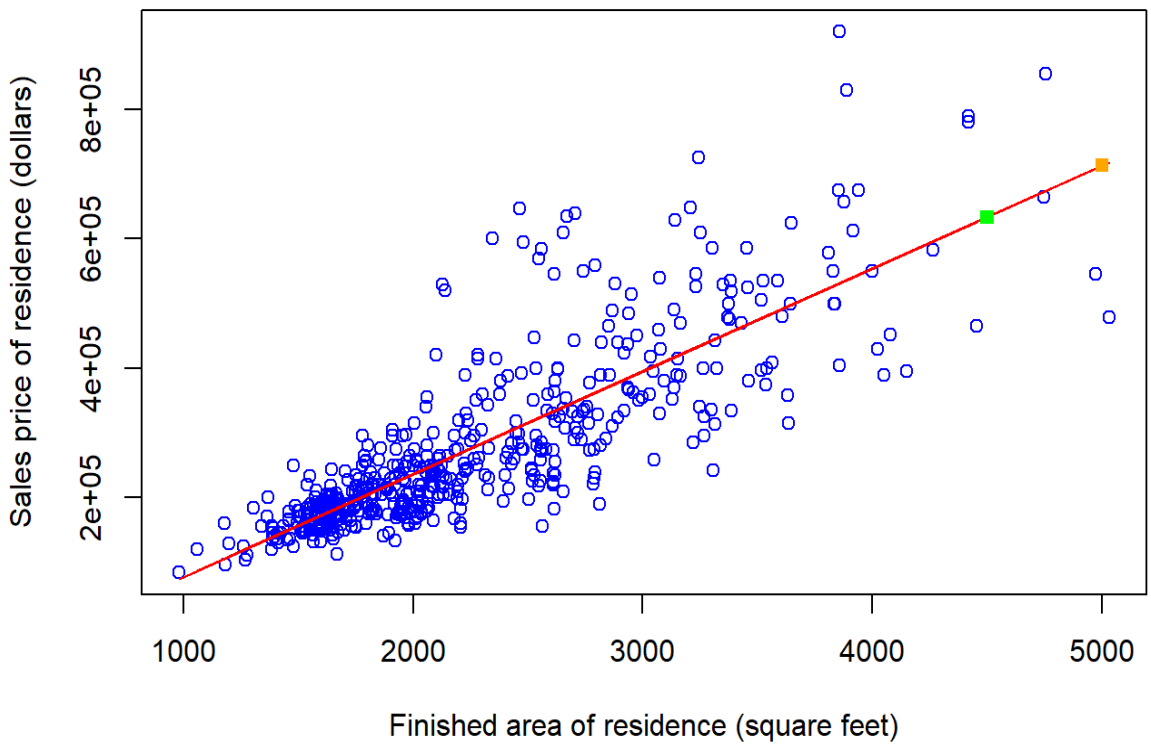

```
rumahku1 <- c(5000)
rumahku1 <- data.frame(rumahku1)
colnames(rumahku1) <- c("X1")
predict(model1, newdata=rumahku1)
```

```
##        1
## 713318.2
```

```
rumahku2 <- c(5000,6000)
rumahku2 <- data.frame(rumahku2)
colnames(rumahku2) <- c("X1")
predict(model1, newdata=rumahku2)
```

```
##        1        2
## 713318.2 872268.4
```

```
plot(rumah$X1, rumah$sales_price,
     xlab="Finished area of residence (square feet)",
     ylab="Sales price of residence (dollars)",
     col="blue")
lines(rumah$X1,fitted(model1),col="red")
points(rumahku2$X1, predict(model1, newdata=rumahku2), col="orange", pch=15)
points(rumahku$X1, predict(model1, newdata=rumahku), col="green", pch=15)
```



Membentuk Model Regresi Linier Berganda

```
#regresi linier berganda
rumah$umur = 2023 - rumah$X7
str(rumah)
```

```
## 'data.frame':    522 obs. of  14 variables:
##  $ ID         : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sales_price: int  360000 340000 250000 205500 275500 248000 229900 150000 195000 160000 ...
##  $ X1         : int  3032 2058 1780 1638 2196 1966 2216 1597 1622 1976 ...
##  $ X2         : int  4 4 4 4 4 4 3 2 3 3 ...
##  $ X3         : int  4 2 3 2 3 3 2 1 2 3 ...
##  $ X4         : int  1 1 1 1 1 1 1 1 1 0 ...
##  $ X5         : int  2 2 2 2 2 5 2 1 2 1 ...
##  $ X6         : int  0 0 0 0 0 1 0 0 0 0 ...
##  $ X7         : int  1972 1976 1980 1963 1968 1972 1972 1955 1975 1918 ...
##  $ X8         : int  2 2 2 2 2 2 2 2 3 3 ...
##  $ X9         : int  1 1 1 1 7 1 7 1 1 1 ...
##  $ X10        : int  22221 22912 21345 17342 21786 18902 18639 22112 14321 32358 ...
##  $ X11        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ umur       : num  51 47 43 60 55 51 51 68 48 105 ...
```

```
View(rumah)


model2 <- lm(sales_price ~ 1 + X1 + umur, data=rumah)
summary(model2)
```

```
##
## Call:
## lm(formula = sales_price ~ 1 + X1 + umur, data = rumah)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -223897  -36225   -7620   24117  389494
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 70812.320  19596.624   3.613 0.000332 ***
## X1            138.339      5.036  27.470  < 2e-16 ***
## umur        -1883.391    203.025  -9.277  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 73350 on 519 degrees of freedom
## Multiple R-squared:  0.7182, Adjusted R-squared:  0.7172
## F-statistic: 661.5 on 2 and 519 DF,  p-value: < 2.2e-16
```

```r
##pendugaan beta
y <- rumah$sales_price
X <- cbind(1,rumah$X1,rumah$umur)
betaduga2 <- solve(t(X)%*%X)%*%t(X)%*%y
betaduga2
```

```
##             [,1]
## [1,] 70812.3203
## [2,]   138.3387
## [3,] -1883.3908
```