Biostatistics Bonus Assignment #5

Marwa Tawfik Badawy

May 13th, 2019

Cholesterol Dataset

Firstly, we will install the tidyverse package.

Installing tidyverse

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyvers
e 1.2.1 ──
```

```
## ✔ ggplot2 3.1.0       ✔ purrr   0.3.2
## ✔ tibble  2.0.1       ✔ dplyr   0.8.0.1
## ✔ tidyr   0.8.3       ✔ stringr 1.3.1
## ✔ readr   1.3.1       ✔ forcats 0.4.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conf
licts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

Secondly, we will load the data itself.

Loading the data

```
cholesterol = read_tsv("cholesterol.tsv")
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   sex = col_character(),
##   age = col_double(),
##   chol = col_double(),
##   BMI = col_double(),
##   TG = col_double(),
##   rs174548 = col_character(),
##   HTN = col_character(),
##   CHD = col_character()
## )
```

```
glimpse(cholesterol)
```

```
## Observations: 400
## Variables: 9
## $ ID       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …
## $ sex      <chr> "F", "F", "M", "M", "F", "F", "M", "M", "M", "M", "M", …
## $ age      <dbl> 74, 51, 64, 34, 52, 39, 79, 38, 52, 58, 43, 64, 38, 63,…
## $ chol     <dbl> 215, 204, 205, 182, 175, 176, 159, 169, 175, 189, 207, …
## $ BMI      <dbl> 26.2, 24.7, 24.2, 23.8, 34.1, 22.7, 22.9, 24.9, 20.4, 2…
## $ TG       <dbl> 367, 150, 213, 111, 328, 53, 274, 137, 125, 209, 122, 1…
## $ rs174548 <chr> "C/G", "G/G", "C/C", "C/G", "C/C", "C/C", "G/G", "C/G",…
## $ HTN      <chr> "Y", "Y", "Y", "Y", "Y", "N", "Y", "N", "N", "Y", "Y", …
## $ CHD      <chr> "Y", "Y", "N", "N", "N", "N", "N", "N", "N", "N", "N", …
```

Adjusting variables

In order to convert the categorical variables from character to factor.

```
cholesterol = cholesterol %>% mutate (sex = factor(sex),
                                      rs174548 = factor(rs174548),
                                      HTN = factor(HTN),
                                      CHD = factor(CHD))
glimpse(cholesterol)
```

```
## Observations: 400
## Variables: 9
## $ ID       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …
## $ sex      <fct> F, F, M, M, F, F, M, M, M, M, M, F, F, M, F, M, M, M, M…
## $ age      <dbl> 74, 51, 64, 34, 52, 39, 79, 38, 52, 58, 43, 64, 38, 63,…
## $ chol     <dbl> 215, 204, 205, 182, 175, 176, 159, 169, 175, 189, 207, …
## $ BMI      <dbl> 26.2, 24.7, 24.2, 23.8, 34.1, 22.7, 22.9, 24.9, 20.4, 2…
## $ TG       <dbl> 367, 150, 213, 111, 328, 53, 274, 137, 125, 209, 122, 1…
## $ rs174548 <fct> C/G, G/G, C/C, C/G, C/C, C/C, G/G, C/G, C/C, C/C, C/G, …
## $ HTN      <fct> Y, Y, Y, Y, Y, N, Y, N, N, Y, Y, Y, Y, Y, N, Y, Y, Y, Y…
## $ CHD      <fct> Y, Y, N, N, N, N, N, N, N, N, N, N, N, N, N, N, Y, N, Y…
```

```
summary(cholesterol)
```

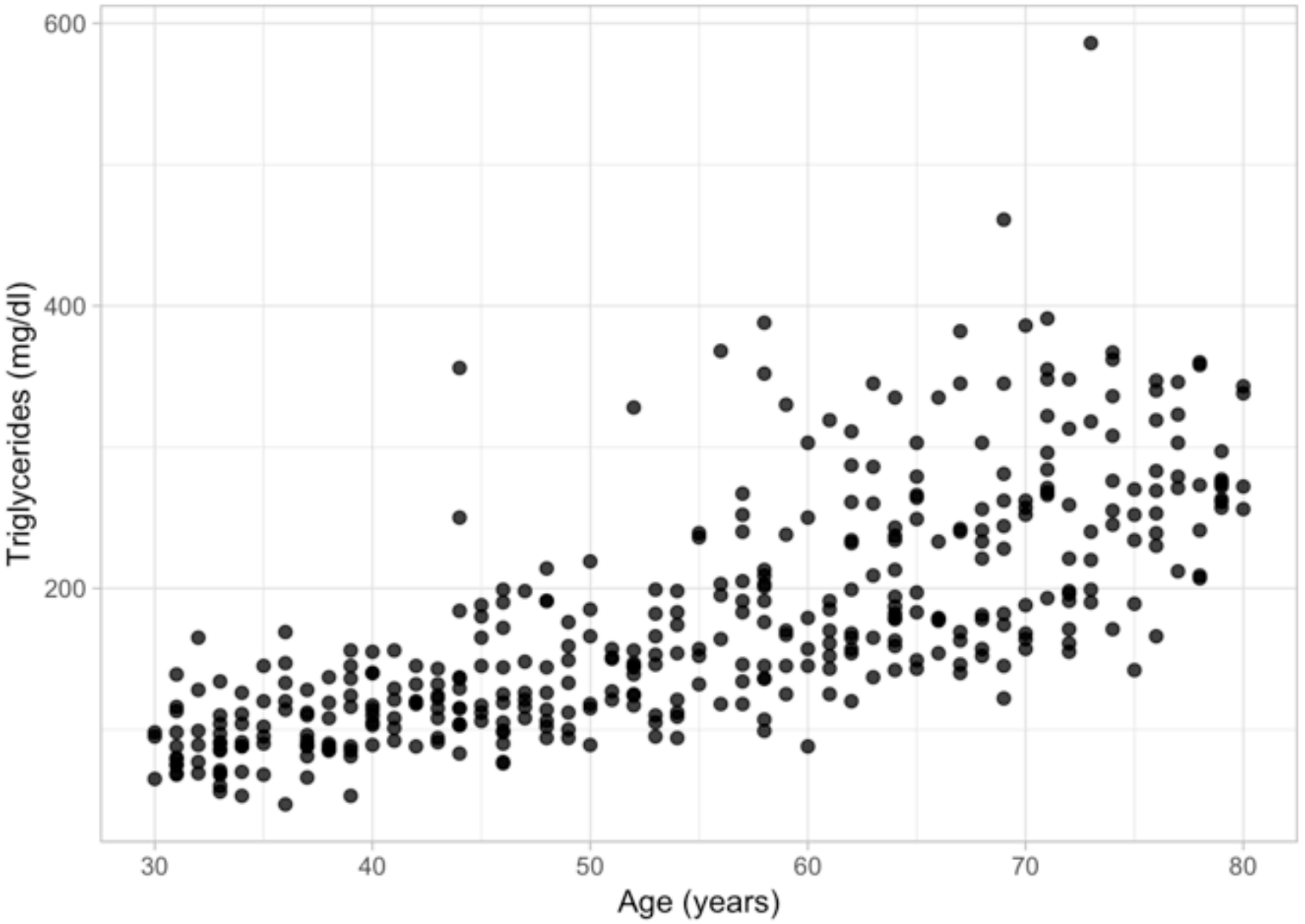
```
##        ID         sex          age              chol             BMI        
##  Min.   :  1.0   F:201   Min.   :30.00   Min.   :117.0   Min.   :19.40  
##  1st Qu.:100.8   M:199   1st Qu.:43.00   1st Qu.:168.0   1st Qu.:22.90  
##  Median :200.5           Median :55.00   Median :184.0   Median :24.60  
##  Mean   :200.5           Mean   :54.82   Mean   :183.9   Mean   :25.00  
##  3rd Qu.:300.2           3rd Qu.:67.00   3rd Qu.:199.2   3rd Qu.:26.73  
##  Max.   :400.0           Max.   :80.00   Max.   :247.0   Max.   :38.80  
##        TG         rs174548  HTN     CHD    
##  Min.   : 47.0   C/C:227   N: 85   N:273  
##  1st Qu.:114.8   C/G:147   Y:315   Y:127  
##  Median :156.5   G/G: 26                  
##  Mean   :177.4                            
##  3rd Qu.:234.0                            
##  Max.   :586.0
```

a) Determine whether the level of Triglycerides (TG) is associated with age

Now, we will determine whether the level of triglycerides (TG) is associated with age or not.

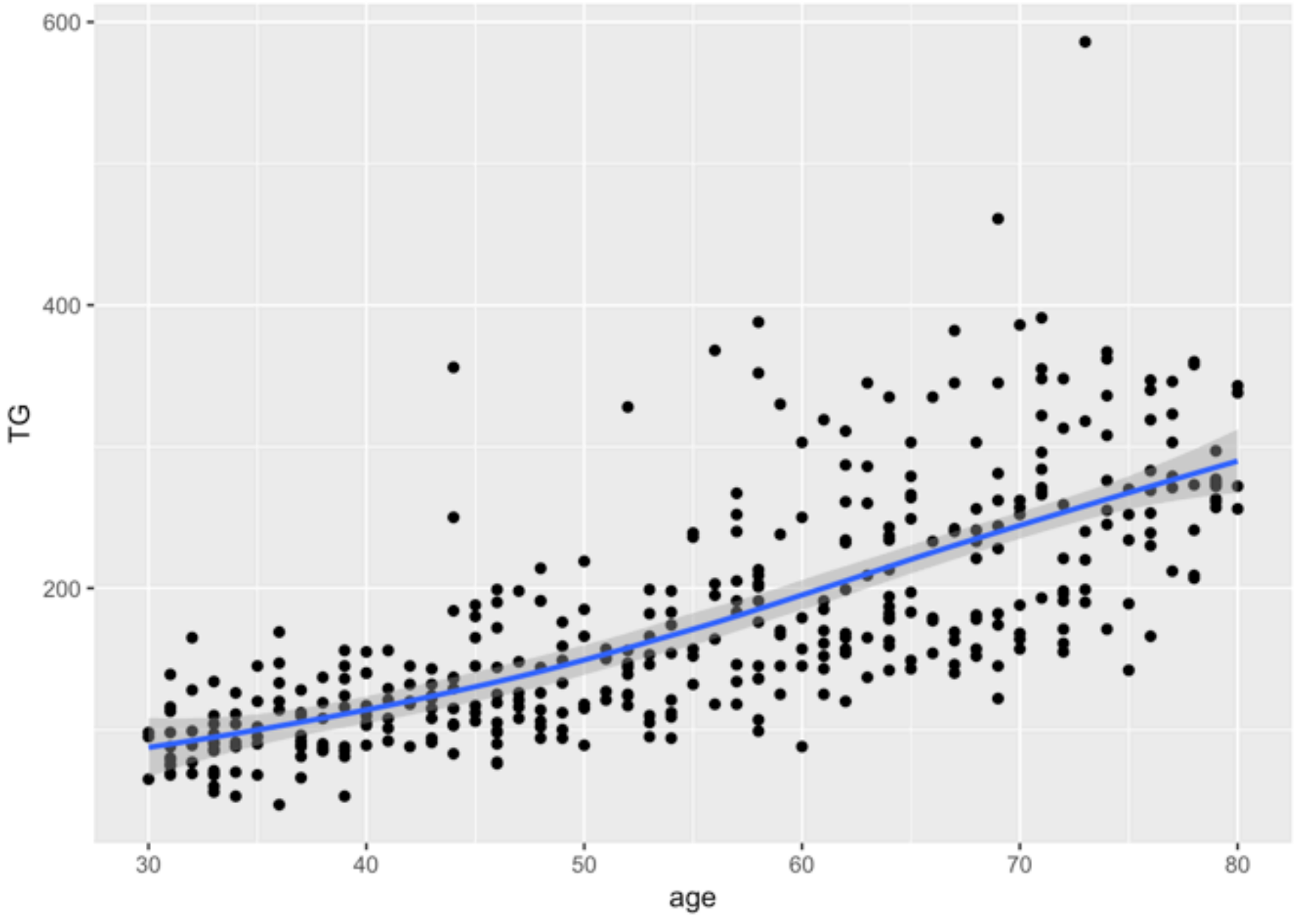
Triglycerides (TG)-age model

```
p = ggplot(cholesterol)
p = p + geom_point(aes(x = age, y = TG), alpha = 0.8, size = 2)
p = p + labs (x = "Age (years)", y = "Triglycerides (mg/dl)")
p = p + theme_light(base_size = 12)
print(p)
```




```
ggplot(data = cholesterol) + geom_point(mapping = aes(x = age, y = TG)) + geom_smo
oth(aes(x = age, y = TG))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



We can interpret that when the age increses, the level of triglycerides is increasing as well. It reaches the levels of 200 to 400 mg/dl from the ages of 60 to 80. The slope here refers to a positive correlation, so triglycerides (TG) is highly associated with the increase in age.

```
model = lm(data = cholesterol, formula = TG ~ age)
broom::tidy(model)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-53.305930	11.1339178	-4.787706	2.383015e-06
age	4.208964	0.1964165	21.428771	2.694609e-68

2 rows

ß0 = -53.3 This means that the estimated average triglycerieds for someone of age = 0 is -53.3

ß1 = 4.2 This means that mean triglycerides is estimated to differ by 4.2 mg/dl for each one year difference in age.
In other words, this will indicate that there's a high association significance

```
summary(model)
```

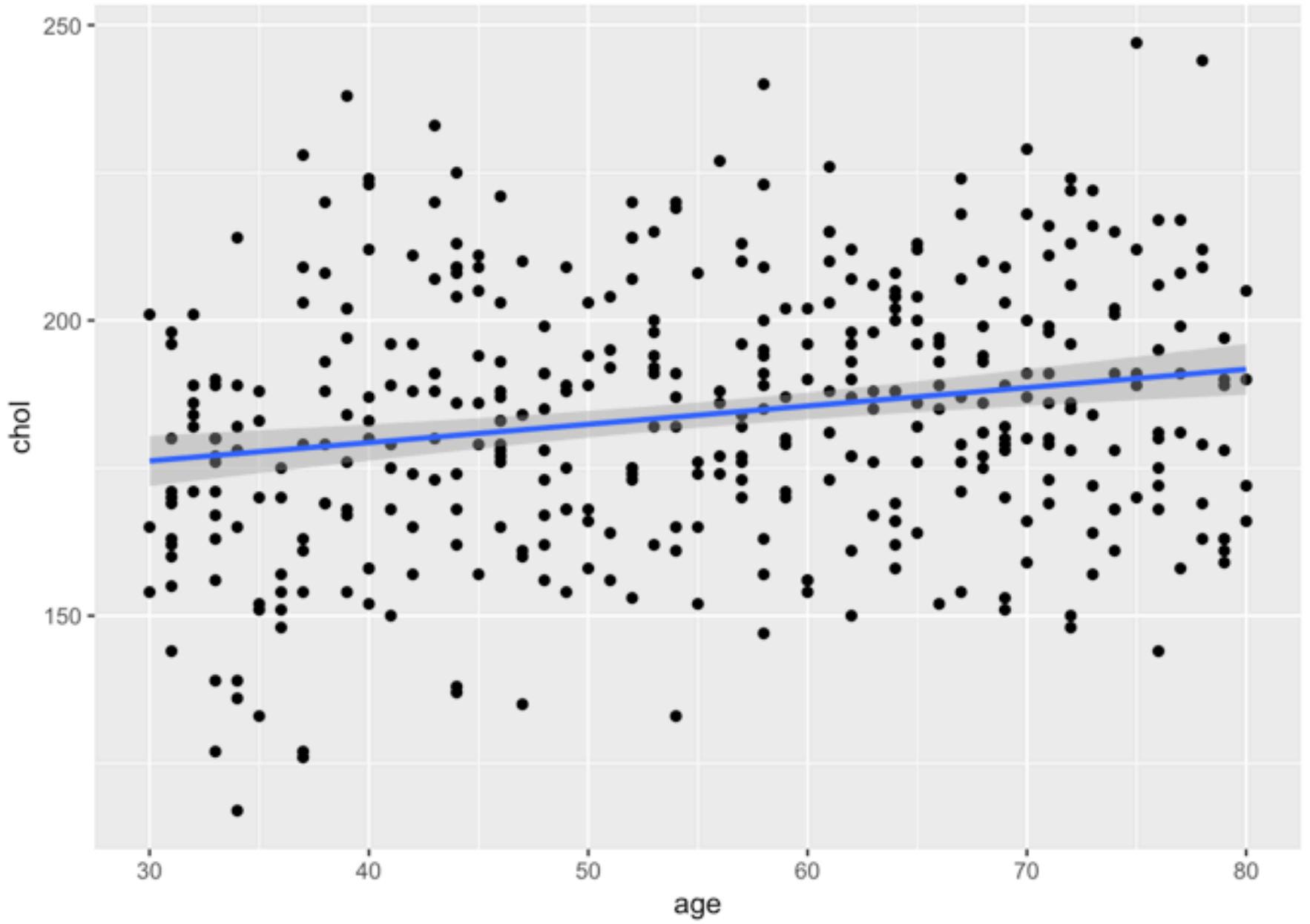
```
##
## Call:
## lm(formula = TG ~ age, data = cholesterol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.37  -36.60   -4.89   24.53  332.05
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
## age           4.2090     0.1964  21.429  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.62 on 398 degrees of freedom
## Multiple R-squared:  0.5357, Adjusted R-squared:  0.5345
## F-statistic: 459.2 on 1 and 398 DF,  p-value: < 2.2e-16
```

b) Compare the cholesterol-age model to the triglycerides-age model

c) Compare the two associations visually

Cholesterol-age model

```
ggplot(data = cholesterol) + geom_point(mapping = aes(x = age, y = chol)) + geom_s
mooth(aes(x = age, y = chol), method = "lm")
```




```
model = lm(data = cholesterol, formula = chol ~ age)
broom::tidy(model)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	166.9016802	4.26488334	39.133938	1.684191e-138
age	0.3103346	0.07523797	4.124707	4.521701e-05

2 rows

ß0 = 166.90 This means that the estimated average serum cholesterol for someone of age = 0 is 166.90

ß1 = 0.31 This means that cholesterol is estimated to differ by 0.31 mg/dl for each one year difference in age.

```
summary(model)
```

```
##
## Call:
## lm(formula = chol ~ age, data = cholesterol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -60.453 -14.643  -0.022  14.659  58.995
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
## age           0.31033    0.07524   4.125 4.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.69 on 398 degrees of freedom
## Multiple R-squared:  0.04099,    Adjusted R-squared:  0.03858
## F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

The cholesterol-age model to the triglycerides-age model comparison.
From the two figures above we can conclude that, triglycerides-age model has a high association significance than cholesterol-age model.