Project 3: Linear Regression

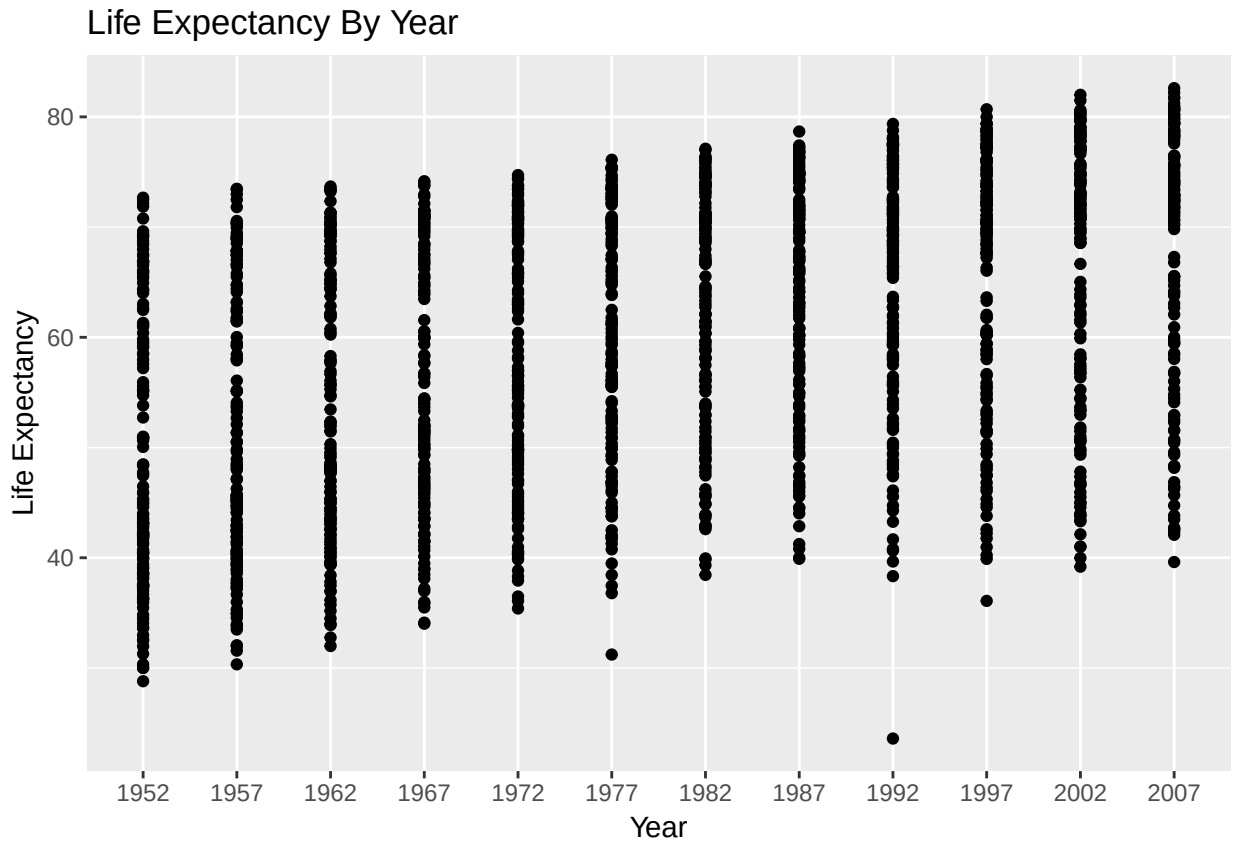
Joseph Kahadze

May 14, 2019

PART 1

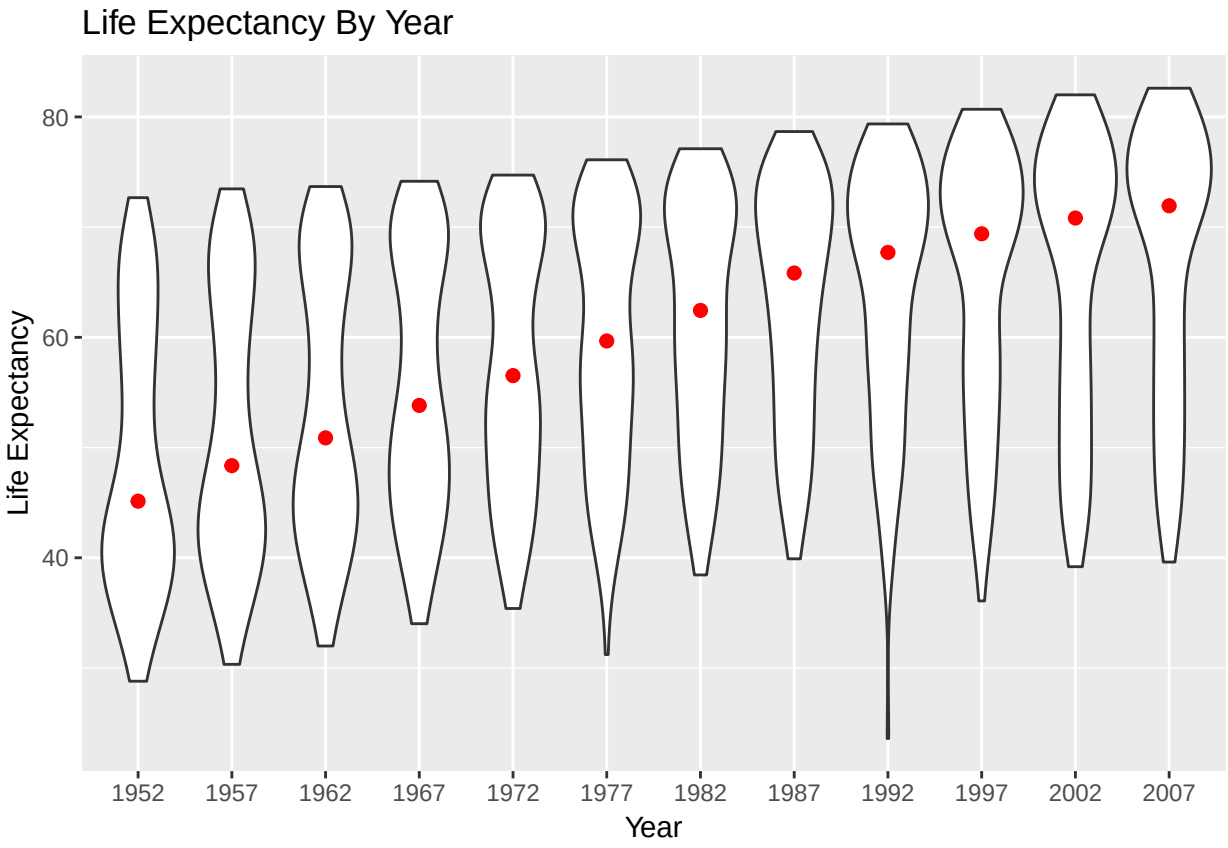
Exercise 1

```
# scatter plot
gapminder %>%
  ggplot(mapping=aes(x=factor(year), y=lifeExp)) +
  geom_point() +
  ggtitle("Life Expectancy By Year") +
  xlab("Year") +
  ylab("Life Expectancy")
```




```
# violin plot
gapminder %>%
  ggplot(mapping=aes(x=factor(year), y=lifeExp)) +
  geom_violin() +
```

```
  ggtitle("Life Expectancy By Year")  +
  stat_summary(fun.y=median, geom="point", size=2, color="red") +
  xlab("Year") + ylab("Life Expectancy")
```



Question 1

Is there a general trend (e.g., increasing or decreasing) for life expectancy across time? Is this trend linear? (answering this qualitatively from the plot, you will do a statistical analysis of this question shortly)

Both the scatter plot and the violin plot show a clear increase in the distribution of life expectancy over the given range of years. The minimum life expectancy rises from approximately 30 years to approximately 40 years, while the maximum life expectancy rises from approximately 70 years to approximately 80 years. The positive trend appears to be linear.

Question 2

How would you describe the distribution of life expectancy across countries for individual years? Is it skewed, or not? Unimodal or not? Symmetric around its center?

The distribution of the violin plot around the median shows that the data is top skewed from 1952 to 1967, has little skew from 1967 to 1977, and it is bottom skewed from 1977 to 2007. The tops of the violin plot are all flat instead of pointed, this suggesting that the distribution is not uni-modal and has multiple maximums. With the exception of 1967 to 1977, the data does not appear to be symmetric around its center. Before 1967 it is bottom heavy while after 1977, it is top heavy.

Question 3

Suppose I fit a linear regression model of life expectancy vs. year (treating it as a continuous variable), and test for a relationship between year and life expectancy, will you reject the null hypothesis of no relationship?

Whether we choose to reject the null hypothesis will depend on the significance level. If we go with the standard 0.05 significance level, my intuition says that there should be sufficient evidence to reject the null hypothesis, and thus claim that there is a positive relationship between year and life expectancy.

Question 4

What would a violin plot of residuals from the linear model in Question 3 vs. year look like?

Since the variance of the distribution for each year appears to be the same, my intuition says that the residuals vs year plot should be fairly horizontal with little positive or negative trend.

Question 5

According to the assumptions of the linear regression model, what should that violin plot look like?

The ideal linear regression model is one which minimizes the residuals thus ideally its residuals vs year plot would be horizontal about 0 on the y-axis.

Exercise 2

```
# linear regression using lm function
modelOne <- lm(lifeExp~year, data=gapminder)

# using tidy function to tidy model stats
tidyOne <- broom::tidy(modelOne)

auto <- modelOne %>%
  augment()

tidyOne
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) -586.      32.3        -18.1 2.90e-67
## 2 year           0.326    0.0163      20.0 7.55e-80
```

Question 6

On average, by how much does life expectancy increase every year around the world?

Since the slope of the linear regression model from exercise 2 is 0.3259, that is the average increase in life expectancy per year across all countries around the world.

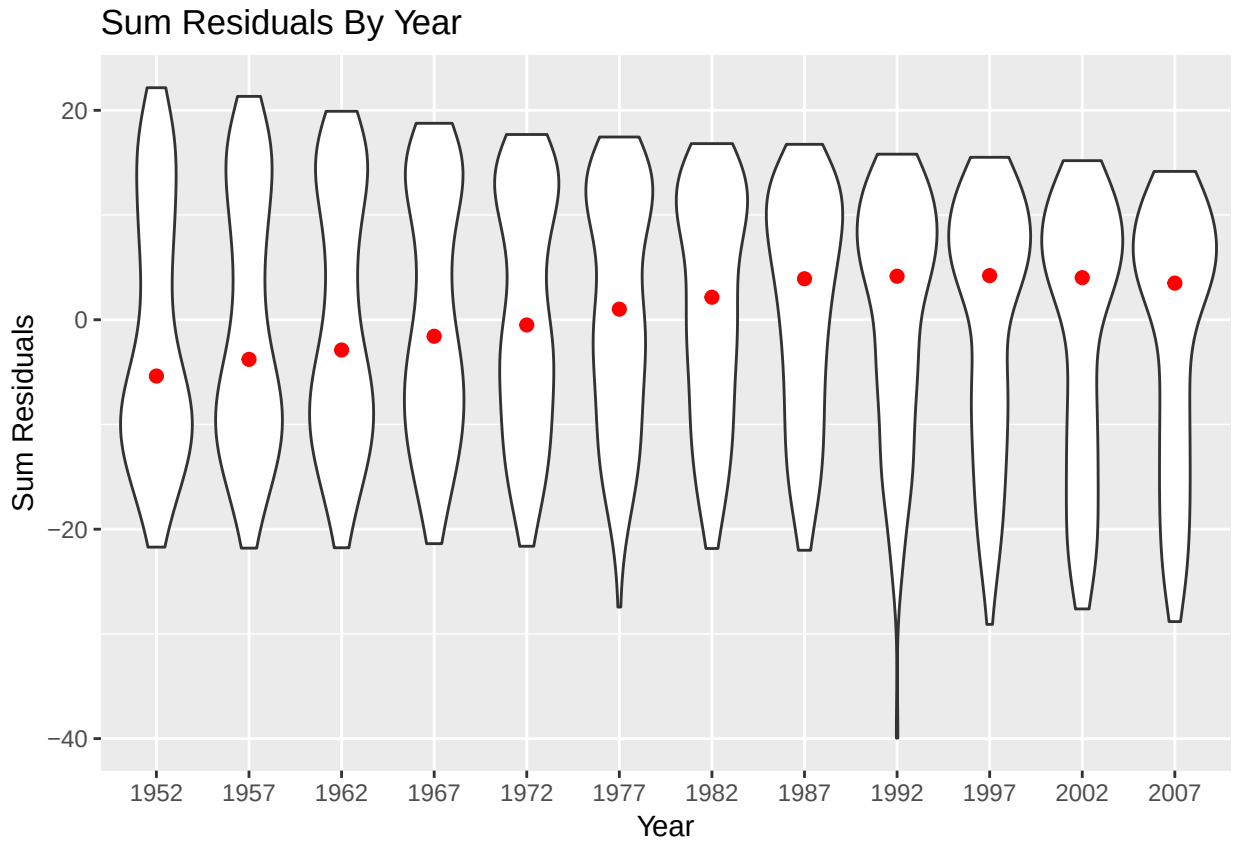
Question 7

Do you reject the null hypothesis of no relationship between year and life expectancy? Why?

Yes there is sufficient evidence to reject the null hypothesis at a significance level of 0.05 because in exercise 2 I get a p-value of 7.546795e-80 (approximately 0), which is less than the significance level. Therefore based on the data there is a positive linear relationship between year and life expectancy in the world.

Exercise 3

```
# create violin plot of sum residuals by year
auto %>% ggplot(mapping=aes(x=factor(year), y=.resid)) + geom_violin() +
  ggtitle("Sum Residuals By Year") +
  stat_summary(fun.y=median, geom="point", size=2, color="red") + xlab("Year") +
  ylab("Sum Residuals")
```



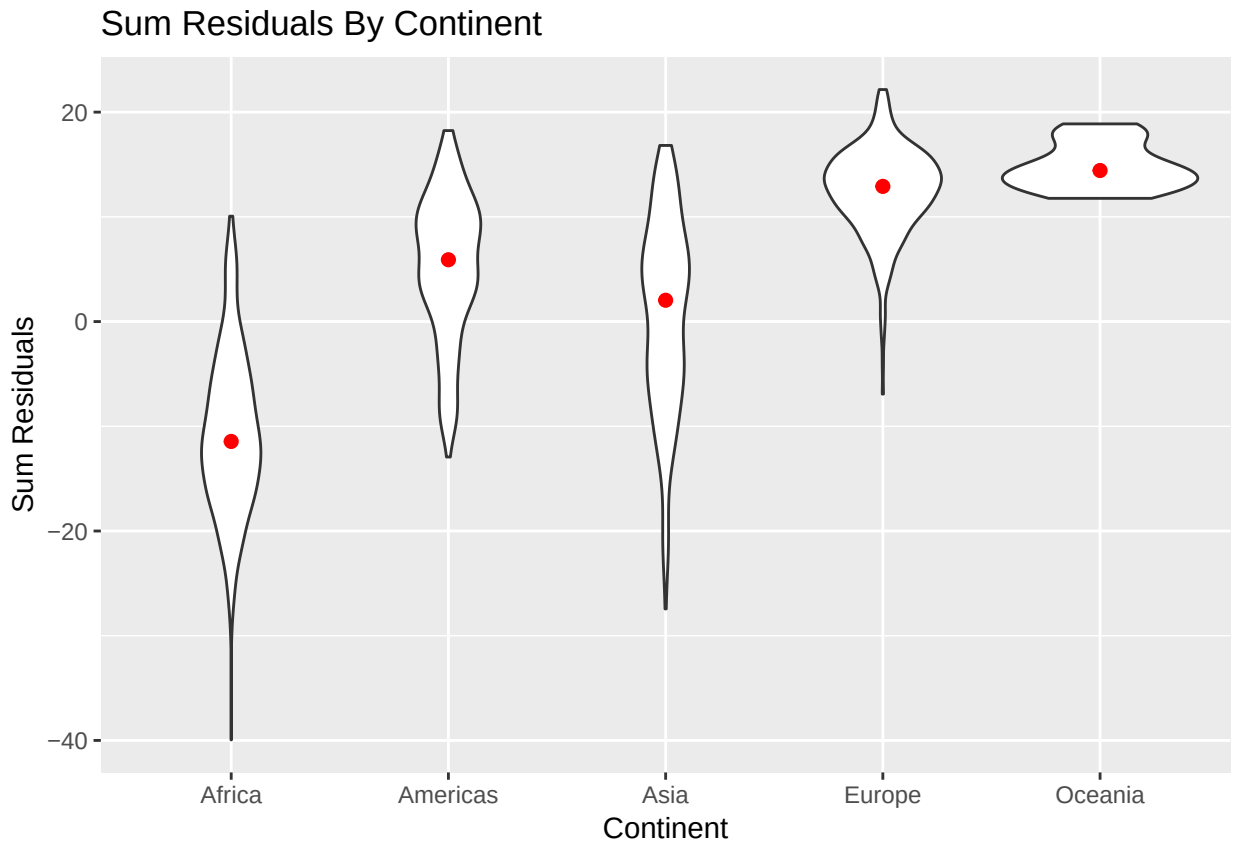
Question 8

Does the plot of Exercise 3 match your expectations?

The plot in exercise 3 mostly matches the expectations I had. The sum residuals vs year plot is mostly centered around 0 as I predicted with a slight negative trend in the max/min residuals and a slight positive trend in the median. Overall, the residuals seem to be minimized around 0 which is what we want from out linear regression model.

Exercise 4

```
# create violin plot of sum residuals by continent
auto %>%
  ggplot(mapping=aes(x=gapminder$continent, y=.resid)) +
  geom_violin() + ggtitle("Sum Residuals By Continent") +
  xlab("Continent") + ylab("Sum Residuals") +
  stat_summary(fun.y=median, geom="point", size=2, color="red")
```



Question 9

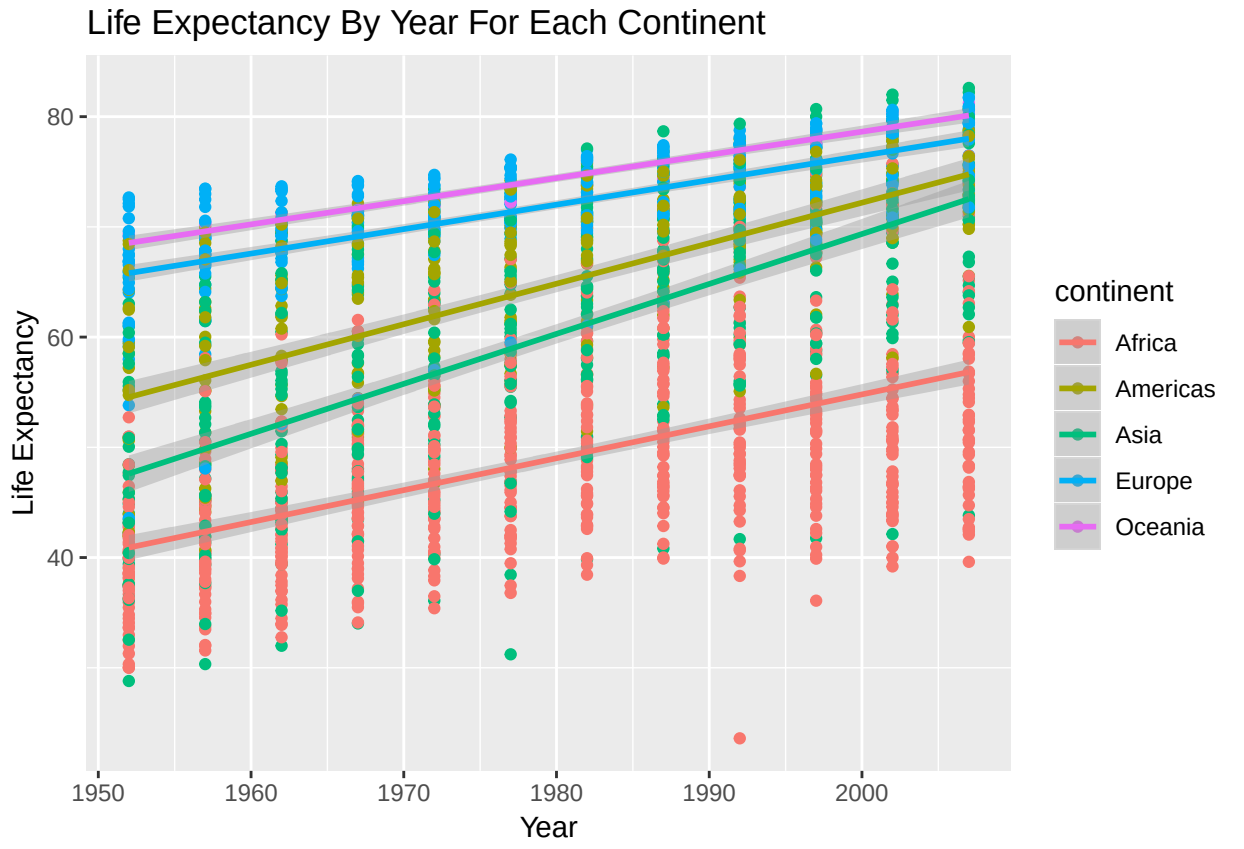
Is there a dependence between model residual and continent? If so, what would that suggest when performing a regression analysis of life expectancy across time?

There appears to be a dependence between sum model residuals and continent as the plots differ between continents. This would suggest that there are factors other than the year at play when performing linear regression analysis, such as the continent.

Exercise 5

```
# plot the life expectancy by year for each continent
gapminder %>%
```

```
ggplot(mapping=aes(x=year, y=lifeExp, group=continent)) +
geom_point(mapping=aes(color=continent)) +
ggtitle("Life Expectancy By Year For Each Continent") +
geom_smooth(method=lm, aes(color=continent)) +
xlab("Year") + ylab("Life Expectancy")
```



Question 10

Based on this plot, should your regression model include an interaction term for continent and year? Why?

Based on this plot it is clear the slope of the line of best fit for year vs life expectancy is different for some continents such as the Americas and Asia, therefore it would be beneficial to include an interaction term for the continent in our model to account for this difference and help minimize the residuals.

Exercise 6

```
# linear regression with added continent interaction term
modelTwo <- lm(lifeExp~year*continent, data=gapminder)
tidyTwo <- broom::tidy(modelTwo)

tidyTwo
```

```
## # A tibble: 10 x 5
##    term                    estimate std.error statistic  p.value
##    <chr>                      <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)             -524.       33.0      -15.9   3.44e-53
##  2 year                       0.290     0.0167    17.4   1.95e-62
##  3 continentAmericas       -139.       57.9       -2.40  1.65e- 2
##  4 continentAsia           -313.       52.9       -5.91  4.14e- 9
##  5 continentEurope          157.       54.5        2.88  4.05e- 3
##  6 continentOceania         182.      171.         1.06  2.87e- 1
##  7 year:continentAmericas     0.0781    0.0292     2.67  7.58e- 3
##  8 year:continentAsia         0.164     0.0267     6.12  1.15e- 9
##  9 year:continentEurope      -0.0676    0.0275    -2.46  1.42e- 2
## 10 year:continentOceania     -0.0793    0.0865    -0.916 3.60e- 1
```

Question 11

Are all parameters in the model significantly different from zero? If not, which are not significantly different from zero?

The parameters closest to 0 are the ones that account for both year and continent, with the largest deviation being .16. The next closest parameter is just the year parameter which has a value of 0.29. Finally, the parameters with just the continents deviate the most from 0, with the largest being Asia with a value of -312.63. In conclusion, this confirms my hypothesis in question 10 that including an interaction term for continent would help minimize the sum residuals for our linear regression model.

Question 12

On average, by how much does life expectancy increase each year for each continent?

```
# get interaction model coeficients from model stat summary
coefs <- tidyTwo[,2] + tidyTwo$estimate[2]
coefs <- coefs[7:10,]
coefs
```

```
## [1] 0.3676509 0.4531224 0.2219321 0.2102724
```

From the results above we can see that life expectancy increases by the following amounts per year per continent from 1952 to 2007:

Africa: **0.28952926 years** Americas: **0.3676509 years** Asia: **0.4531224 years** Europe: **0.2219321 years** Oceania: **0.2102724 years**

Exercise 7

```
# use anova() function to perform anova testing on our two models
anova(modelOne)
```

```
## Analysis of Variance Table
##
## Response: lifeExp
```

```
##             Df Sum Sq Mean Sq F value    Pr(>F)
## year         1  53919   53919   398.6 < 2.2e-16 ***
## Residuals 1702 230229     135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(modelTwo)
```

```
## Analysis of Variance Table
##
## Response: lifeExp
##                   Df Sum Sq Mean Sq  F value    Pr(>F)
## year               1  53919   53919 1046.028 < 2.2e-16 ***
## continent          4 139343   34836  675.812 < 2.2e-16 ***
## year:continent     4   3566     892   17.296 6.463e-14 ***
## Residuals       1694  87320      52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 13

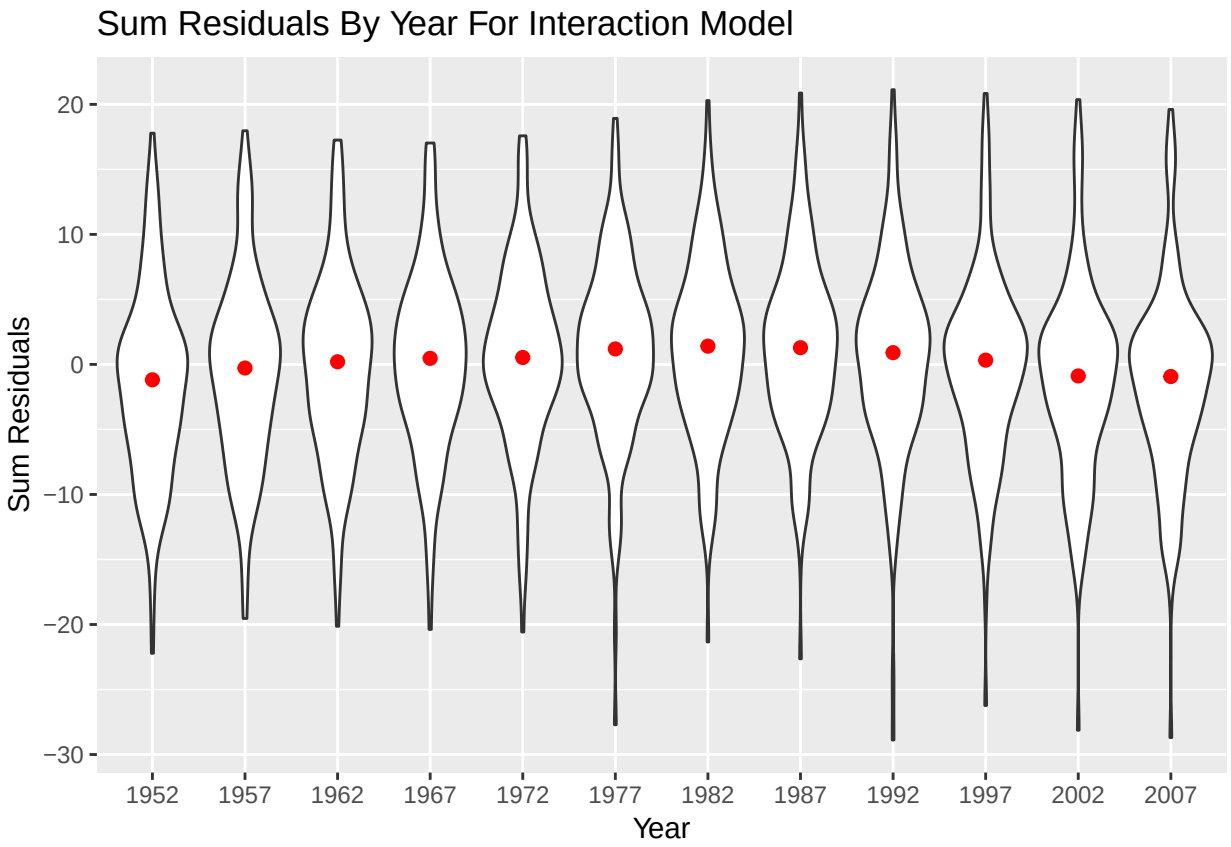
Is the interaction model significantly better than the year-only model? Why?

Comparing the sum residuals from exercise 7 for just the year linear model with the residuals for the combined year and continent linear model, we can see that the sum residuals decreased from 1702 to 1694. Since the interaction model minimizes residuals better than the year-only model, we have sufficient evidence to conclude that the interaction model is better than the year-only model.

Exercise 8

```
auto2 <- modelTwo %>%
 augment()

# plot sum residuals of interaction model
auto2 %>% ggplot(mapping=aes(x=factor(year), y=.resid)) + geom_violin() +
  ggtitle("Sum Residuals By Year For Interaction Model") +
  stat_summary(fun.y=median, geom="point", size=2, color="red") +
  xlab("Year") + ylab("Sum Residuals")
```


Sum Residuals By Year For Interaction Model
Sum Residuals
20
10
0
−10
−20
−30
1952
1957
1962
1967
1972
1977
1982
1987
1992
1997
2002
2007
Year