Regression

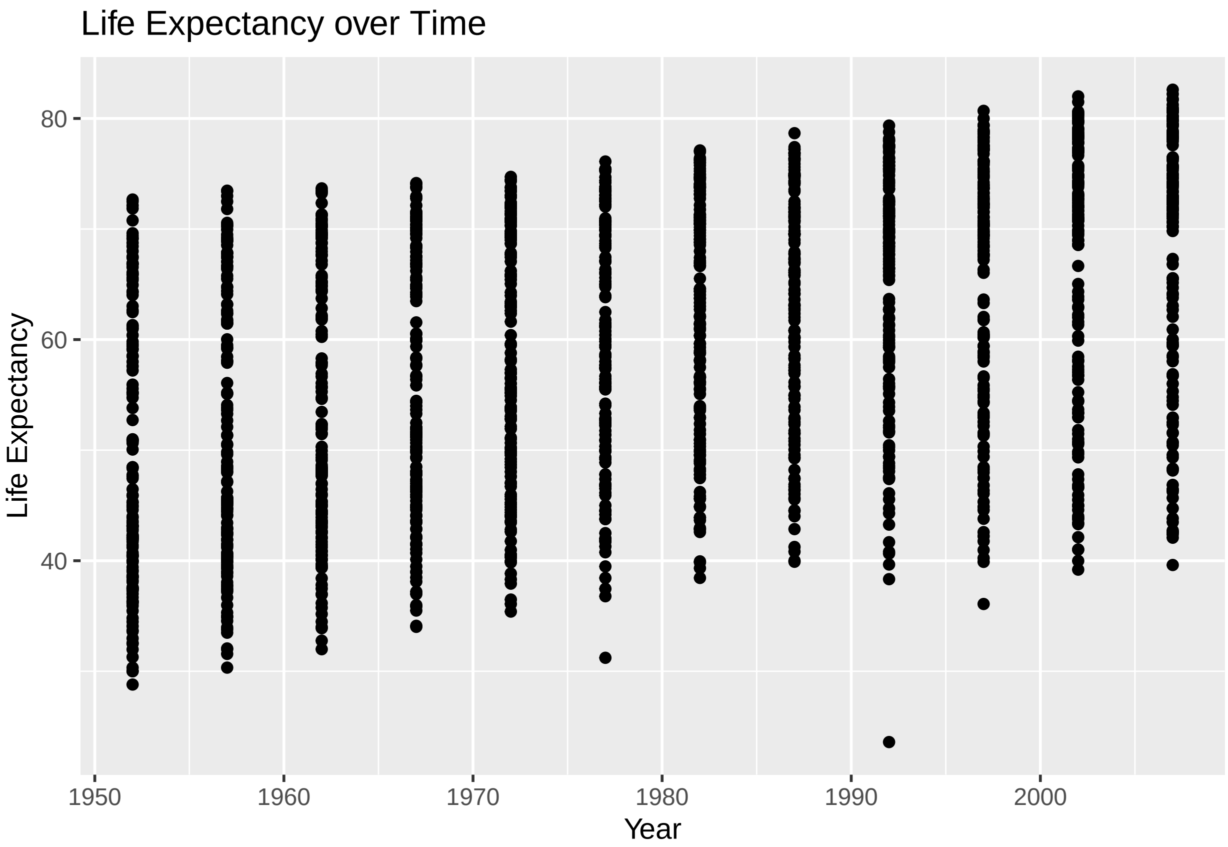
Zackary Frazier

4/13/2020

Exercise 1: Make a scatter plot of life expectancy across time

```
expected <- mean(gapminder$lifeExp)

gapminder %>%
  ggplot(aes(x = year, y = lifeExp)) +
  geom_point() +
  labs(title="Life Expectancy over Time", x="Year", y="Life Expectancy")
```

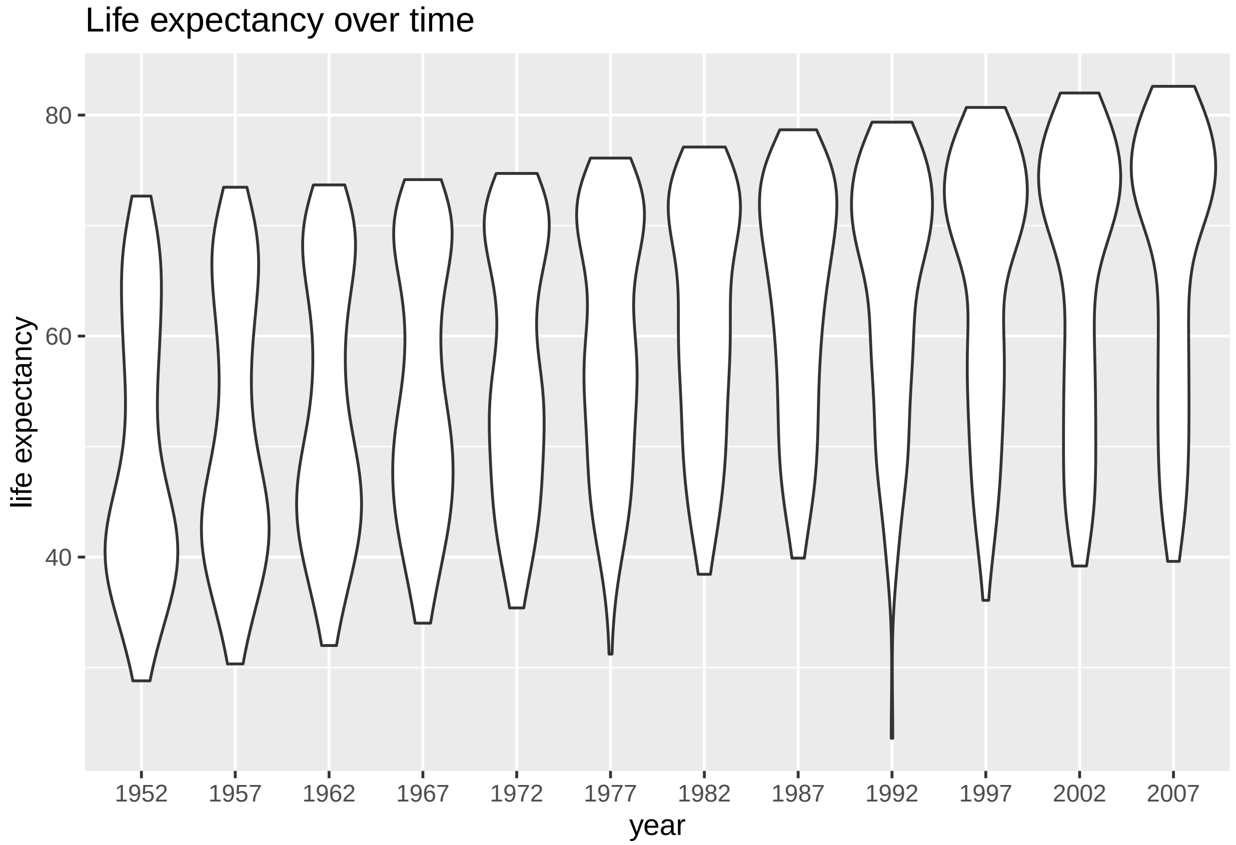


# Question 1: Is there a general trend (e.g., increasing or decreasing) for life expectancy across time? Is this trend linear? (answering this qualitatively from the plot, you will do a statistical analysis of this question shortly)

It seems the general trend is an increasing life expectancy over time. From 1950 to 2000 the max life expectancies and smallest life expectancies all increased by about 20 years.

A slightly different way of making the same plot is looking at the distribution of life expectancy across countries as it changes over time:

```
gapminder %>%
  ggplot(aes(x=factor(year), y=(lifeExp))) +
    geom_violin() +
    labs(title="Life expectancy over time",
         x = "year",
         y = "life expectancy")
```



This type of plot is called a violin plot, and it displays the distribution of the variable in the y-axis for each value of the variable in the x-axis.

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?

It appears that it began slightly bimodal, with one focus around the bottom of the distribution, and another at the top. This probably reflects the higher levels inequality that were experienced throughout most of the 20th century. As the years enter the 21st century the lower end of the distribution expands a bit, but the graph generally moves toward a unimodal distribution centered at the top (at around 70 year life expectancies). These distributions are certainly not symmetric. They begin heavily skewed toward the lower end of life expectancy, and over time the skew gradually moves toward the higher end of life expectancy.

Based on this plot, consider the following questions.

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? (do this without fitting the model yet. I am testing your intuition.)

Yes. I would reject that hypothesis. The scatter plot and the violin plots clearly show some 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? (Again, don't do the analysis yet, answer this intuitively)

It would look like an a bunch of evenly distributed violin plots with the residuals evenly scattered around x=0.

Question 5: According to the assumptions of the linear regression model, what should that violin plot look like?

Ideally the residuals should appear to be evenly scattered around y=0. with roughly constant variance on both sides.

Exercise 2: Fit a linear regression model using the lm function for life expectancy vs. year (as a continuous variable). Use the broom::tidy to look at the resulting model.

```r
library(broom)

regression <- lm(lifeExp ~ year, data=gapminder)
tidy(regression)
```

```
## # 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?

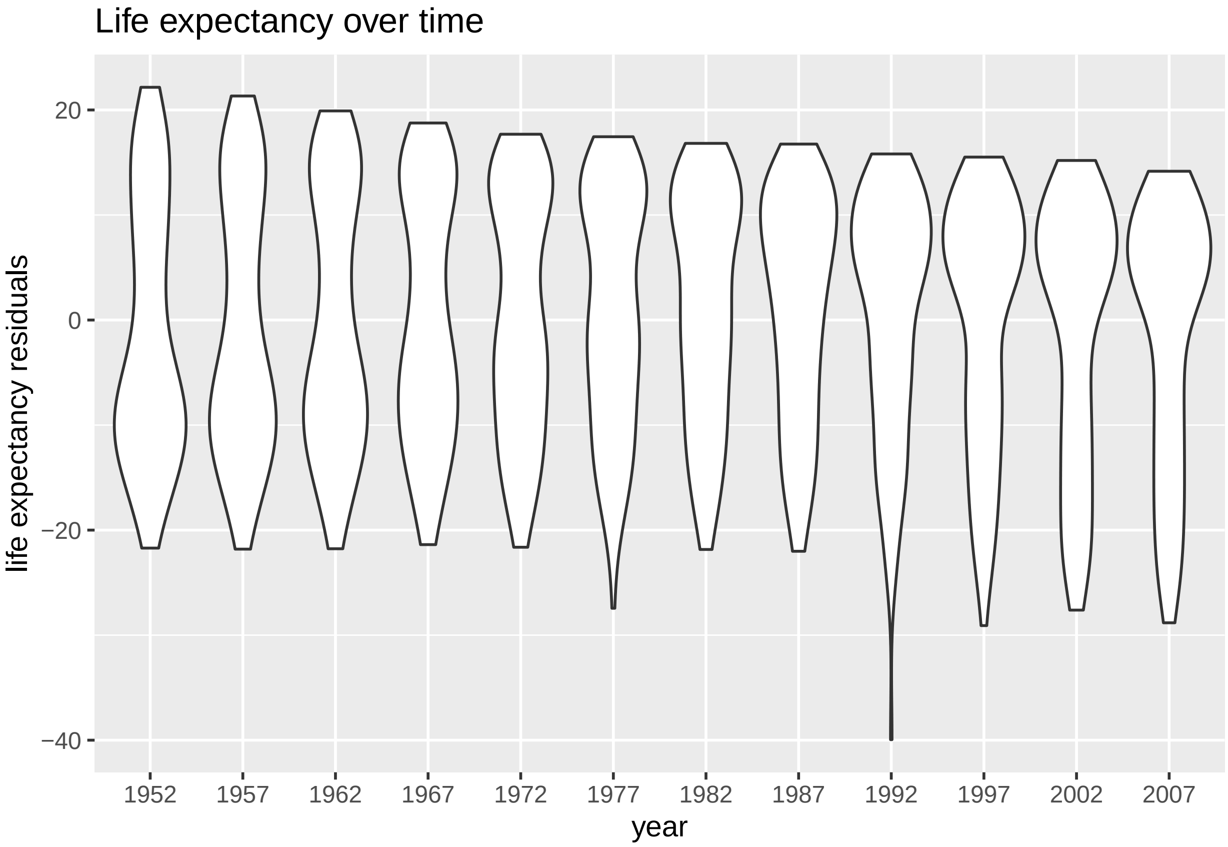
On average, the life expectancy increases by 0.33 years per year.

Question 7: Do you reject the null hypothesis of no relationship between year and life expectancy? Why?

Yes. Our p-value is tiny. If we use the same threshold as the last project of $\alpha = 0.05$, our p-value is well below α.

Exercise 3: Make a violin plot of residuals vs. year for the linear model from Exercise 2 (use the broom::augment function).

```r
augment(regression) %>%
  ggplot(aes(x=factor(year), y=.resid)) +
  geom_violin() +
  labs(title="Life expectancy over time",
     x = "year",
     y = "life expectancy residuals")
```

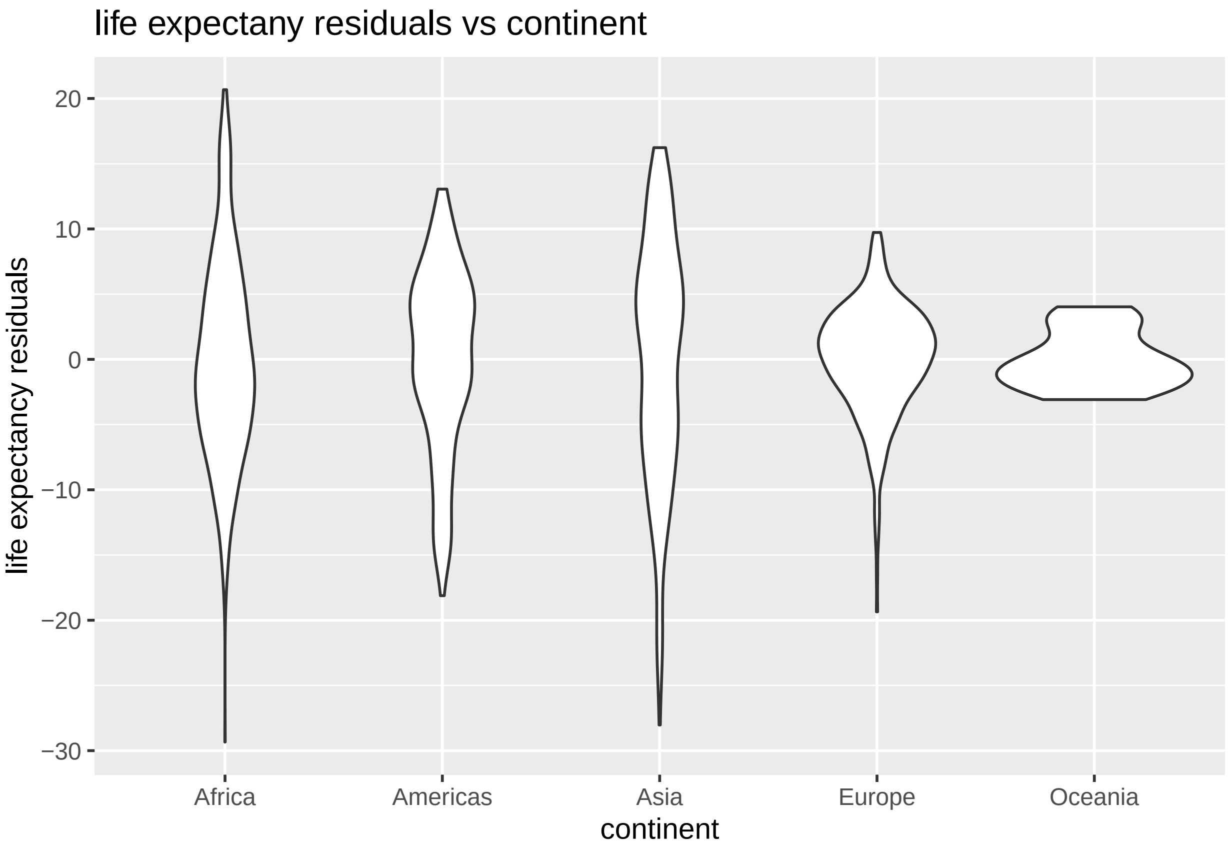


Question 8: Does the plot of Exercise 3 match your expectations (as you answered Question 4)?

No. But I was thinking in terms of distance from the mean, this calculation for the residuals appears to have been constructed via distance from the regression line. The most notable difference I see is that the inaccuracy of the model stays about the same throughout the plotting, but it shifts from overestimating to underestimating around the end.

Exercise 4: Make a boxplot (or violin plot) of model residuals vs. continent.

```r
regression <- lm(lifeExp ~ year + continent, data=gapminder)
augment(regression) %>%
  ggplot(aes(x=continent, y=.resid)) +
  geom_violin() +
  labs(title="life expectany residuals vs continent",
     x = "continent",
     y = "life expectancy residuals")
```

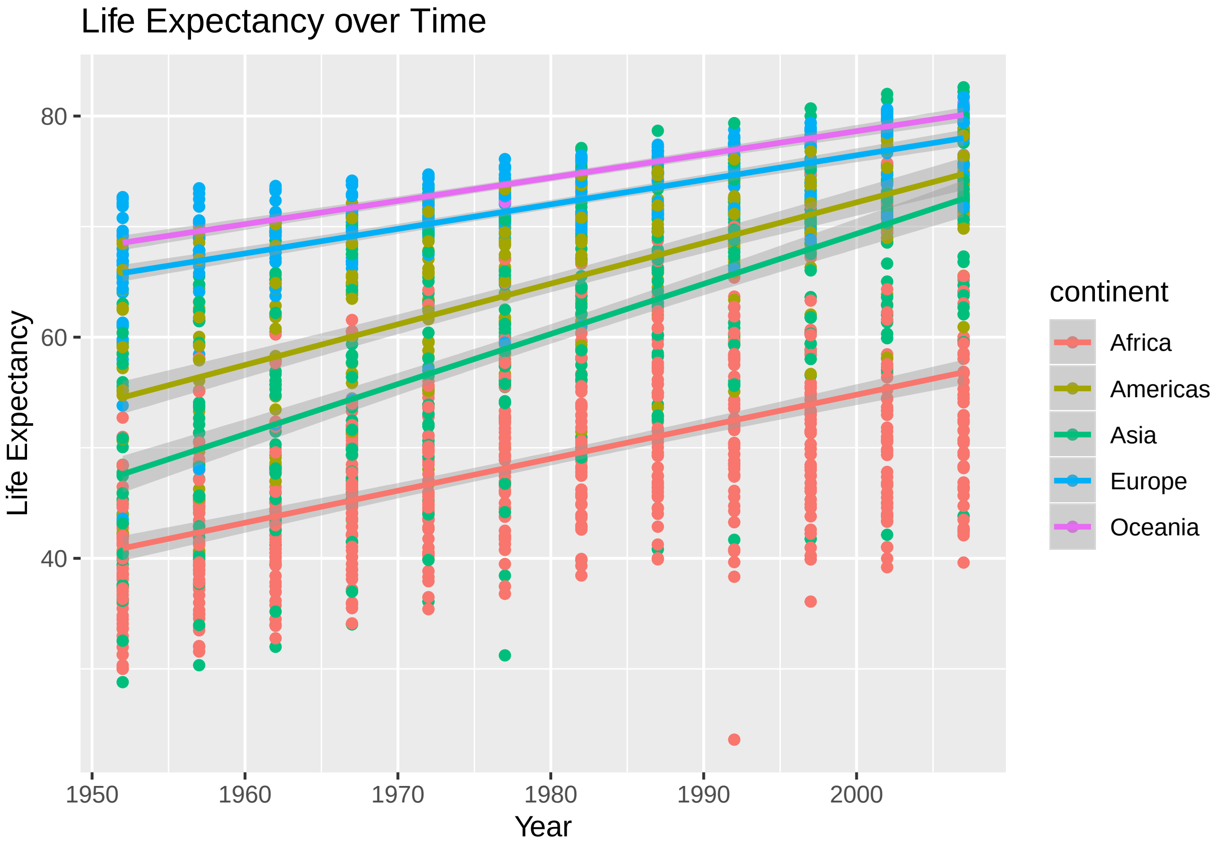


Question 9: Is there a dependence between model residual and continent? If so, what would that suggest when performing a regression analysis of life expectancy across time?

It appears so, yes. It seems that model residuals are more accurate for people in Oceania and Europe, and can be all over the place for people born in Asia, Africa and the Americas, with Africa and Asia having the highest variances.

Exercise 5: Use geom_smooth(method=lm) in ggplot as part of a scatter plot of life expectancy vs. year, grouped by continent (e.g., using the color aesthetic mapping).

```
gapminder %>%
  ggplot(aes(x = year, y = lifeExp, color=continent)) +
  geom_point() +
  labs(title="Life Expectancy over Time", x="Year", y="Life Expectancy") +
  geom_smooth(method="lm")
```



Question 10: Based on this plot, should your regression model include an interaction term for continent and year? Why?

Yes. The average life expectencies increase at different rates for each respective continent. The life expectencies of people from different continents also start from different starting points and end at different points, most dramatically Africa. It is clear from this graph that one's life expectancy can be modelled as a function of one's continent.

Exercise 6: Fit a linear regression model for life expectancy including a term for an interaction between continent and year. Use the broom::tidy function to show the resulting model.

```
regression <- lm(lifeExp ~ year * continent, data=gapminder)
tidy(regression)
```

```
## # 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?

No. The paramater for the effect of year on life expectancy in Africa is roughly 0.29, stating that every year the average life expectancy across Africa increases by 0.29 years per year. Also the the parameter for the effect of year on life expectancy in Asia is substantially higher than 0. This paramter having a value of 0.16 implies the life expectancy in Asia is increasing substantially faster than the life expectancies on other continents.

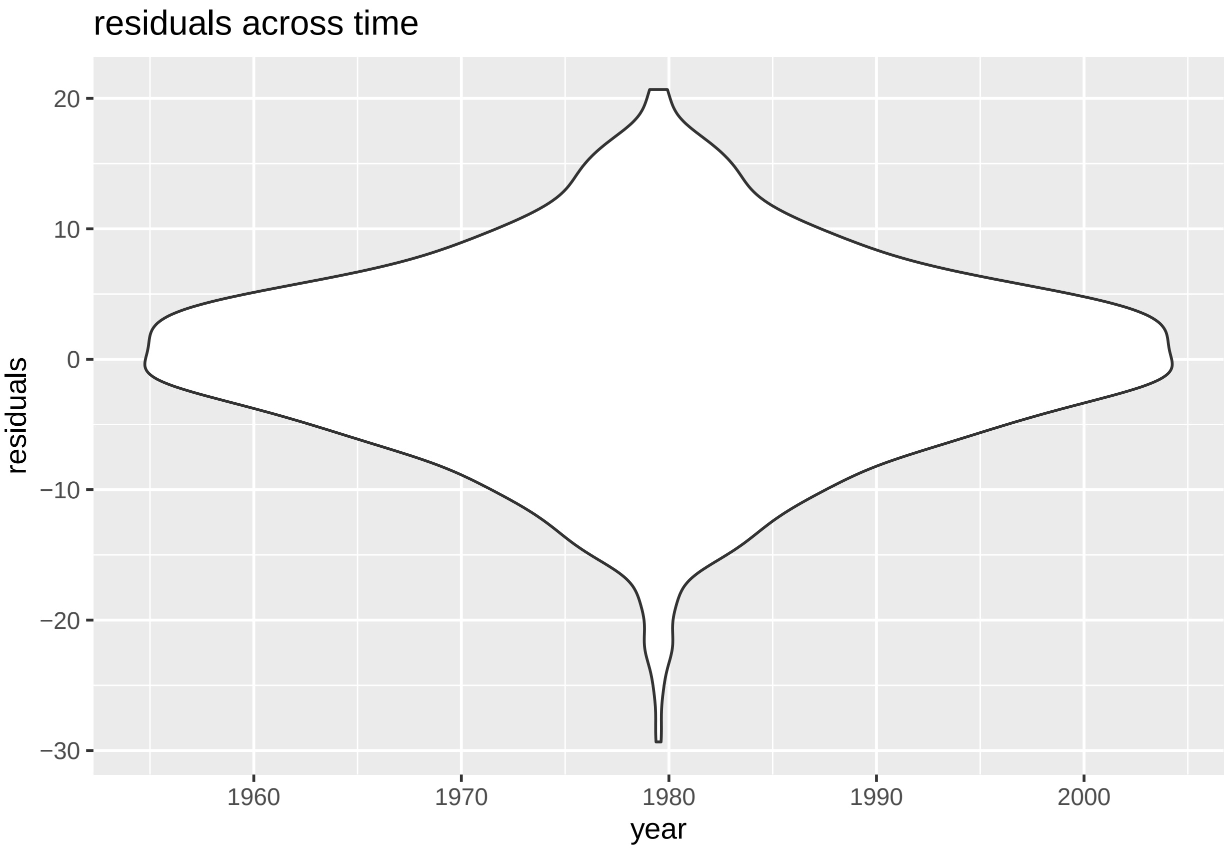
Question 12: On average, by how much does life expectancy increase each year for each continent? (Provide code to answer this question by extracting relevant estimates from model fit)

```
regression <- lm(lifeExp ~ year*continent, data=gapminder)
tidy(regression) %>%
  filter(-2 < estimate & estimate < 2 ) %>%
  mutate(term = ifelse(term == 'year', 'year:continentAfrica', term)) %>%
  select(term, estimate)
```

```
## # A tibble: 5 x 2
##   term                   estimate
##   <chr>                     <dbl>
## 1 year:continentAfrica     0.290
## 2 year:continentAmericas   0.0781
## 3 year:continentAsia       0.164
## 4 year:continentEurope    -0.0676
## 5 year:continentOceania   -0.0793
```

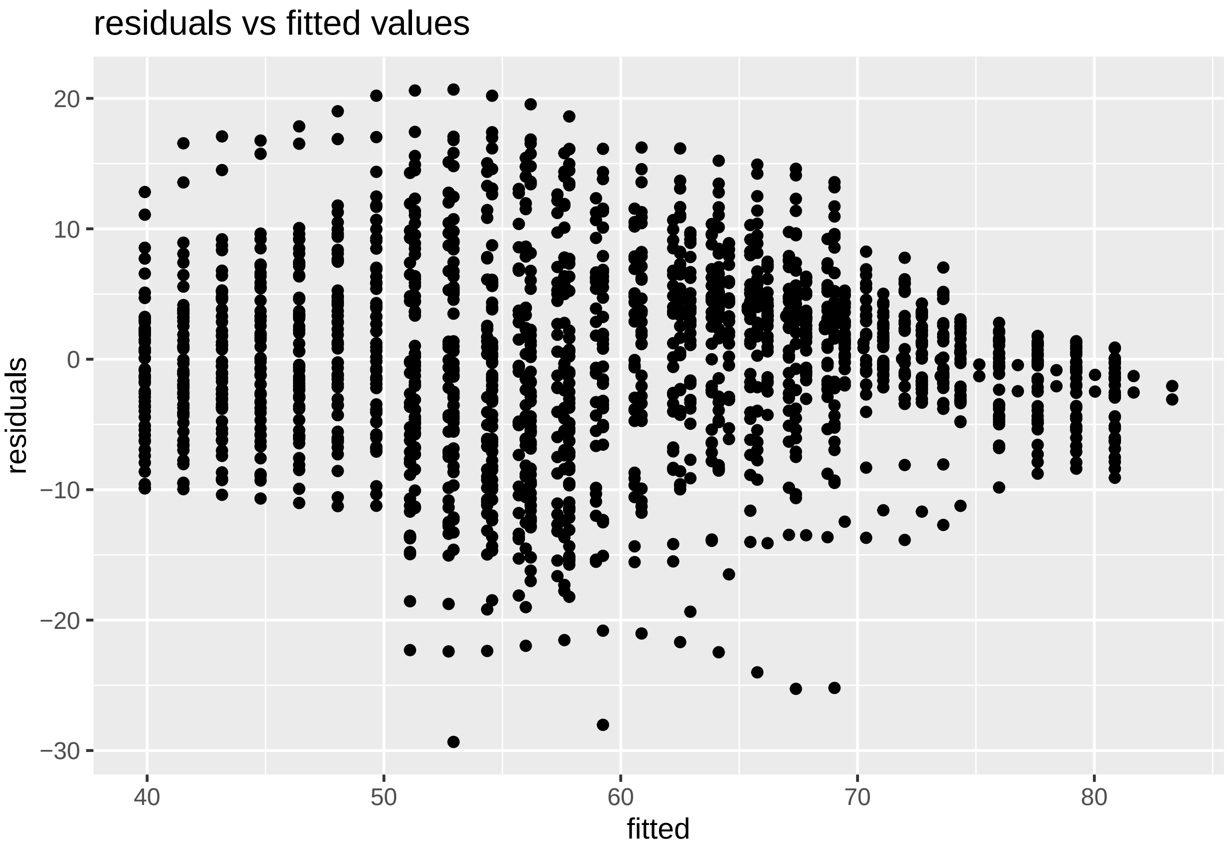
Exercise 7: Make a residuals vs. year violin plot for the interaction model. Comment on how well it matches assumptions of the linear regression model. Do the same for a residuals vs. fitted values model.

```
regression <- lm(lifeExp ~ year + continent, data=gapminder)
augment(regression) %>%
  ggplot(aes(x=year, y=.resid)) +
  geom_violin() +
  labs(title="residuals across time",
     x = "year",
     y = "residuals")
```



The general symmetry of this violin plot suggests that our variance is constant. This also suggest that our residuals are indeed independent and identically distributed.

```
regression <- lm(lifeExp ~ year + continent, data=gapminder)
augment(regression) %>%
  ggplot(aes(x=.fitted, y=.resid)) +
  geom_point() +
  labs(title="residuals vs fitted values",
     x = "fitted",
     y = "residuals")
```



This plot is centered around 0. This implies our inital assumption that the data is a linear distribution is also true.