Regression Analysis of Gapminder Data

Daniel Brewer

5/4/2019

Q1: There is a general trend when creating a scatter plot of life expectancy across time. Using qualitative judgement, the trend is that life expectancy is linearly increasing across time.

```
library(gapminder)
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------- tidyverse 1.2.1 --
```

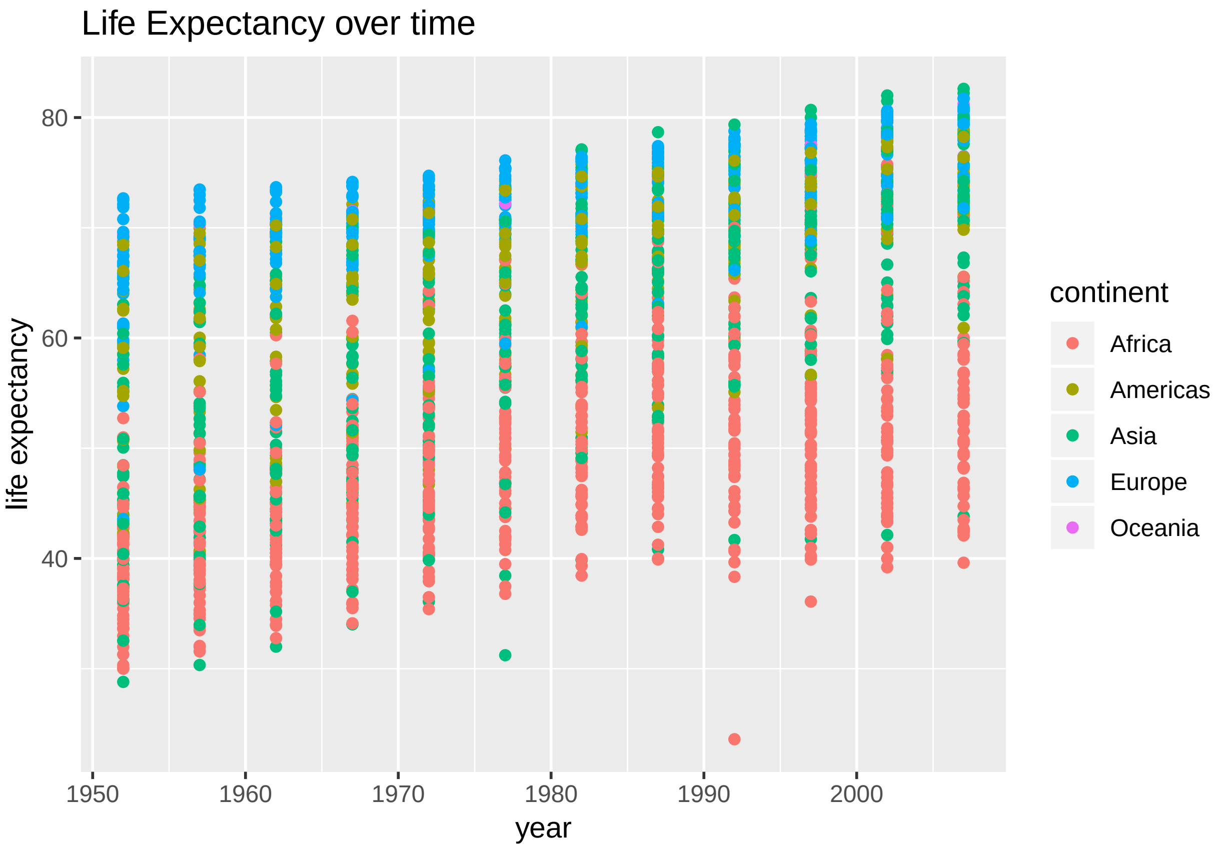
```
## v ggplot2 3.1.1       v purrr   0.3.2
## v tibble  2.1.1       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v ggplot2 3.1.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(broom)
data(gapminder)
gapminder
```

```
## # A tibble: 1,704 x 6
##    country     continent  year lifeExp      pop gdpPercap
##    <fct>       <fct>     <int>   <dbl>    <int>     <dbl>
##  1 Afghanistan Asia       1952    28.8  8425333      779.
##  2 Afghanistan Asia       1957    30.3  9240934      821.
##  3 Afghanistan Asia       1962    32.0 10267083      853.
##  4 Afghanistan Asia       1967    34.0 11537966      836.
##  5 Afghanistan Asia       1972    36.1 13079460      740.
##  6 Afghanistan Asia       1977    38.4 14880372      786.
##  7 Afghanistan Asia       1982    39.9 12881816      978.
##  8 Afghanistan Asia       1987    40.8 13867957      852.
##  9 Afghanistan Asia       1992    41.7 16317921      649.
## 10 Afghanistan Asia       1997    41.8 22227415      635.
## # ... with 1,694 more rows
```

```
life_plot <- gapminder %>% ggplot(mapping=aes(x=year, y=lifeExp)) + geom_point(mapping=aes(color=contin

life_plot
```



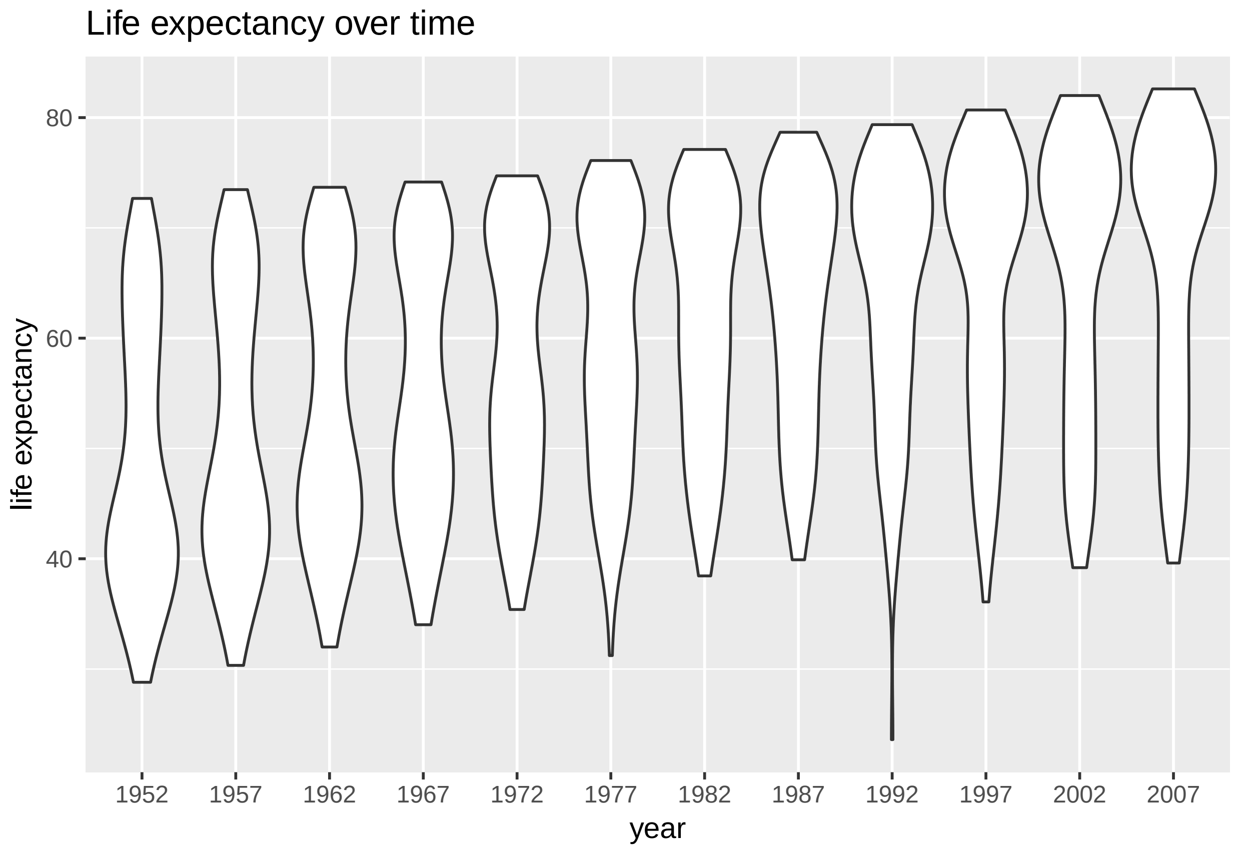
Q2: Using a violin plot shows that the distribution of life expectancy across countries begins as more concentrated below the median, but as years goes on it becomes more concentrated above the median. This shows that the data is skewed and not symmetric around its center, especially at the start and end years. Overall, the distributions each year are unimodal, but between 1957-1967 the distribution takes on a bimodal shape.

Q3: If this model were to be fit, I would reject the null hypothesis of no relationship.

Q4: A violin plot of residuals from the linear model in Q3 vs. year would look very similar to the plot below (due to the difference between predicted and actual value), except the trend would be downwards (due to the increasing linear relationship in the model below). If linear, the residuals will be centered around 0 similarly to the values being centered around 60 in the plot below.

Q5: If the assumption is that there is no relationsip, we would see a plot very different from the one below, with residual values not centered around 0. This would indicate there is no relationship due to a variation in predicted vs actual distance.

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



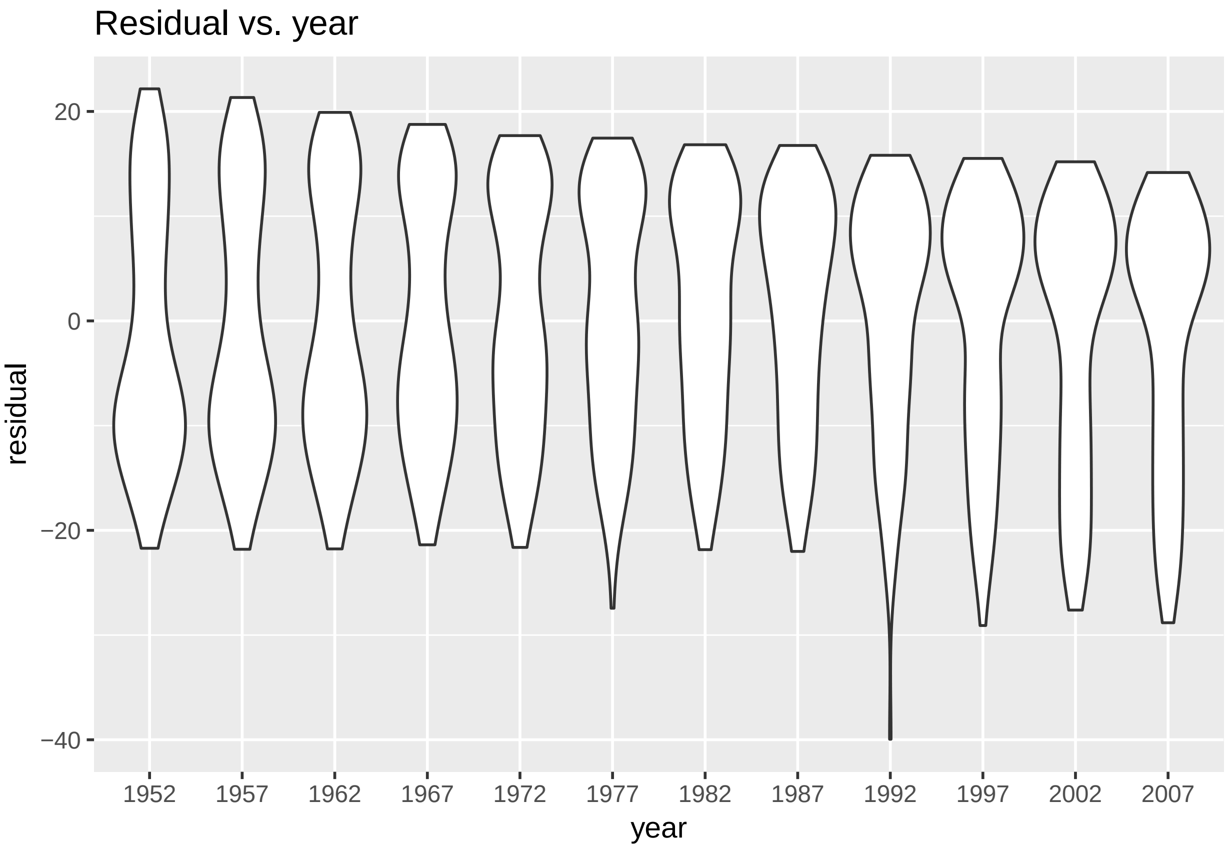
Q6: On average, life expectancy increases by 0.3259 per year around the world.
Q7: Due to the low p value (< 0.05) seen below, the null hypothesis of no relationship between year and life expectancy can be rejected.

```
linear_co <- lm(lifeExp~year, data=gapminder)
linear_co %>% tidy()
```

```
## # 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
```

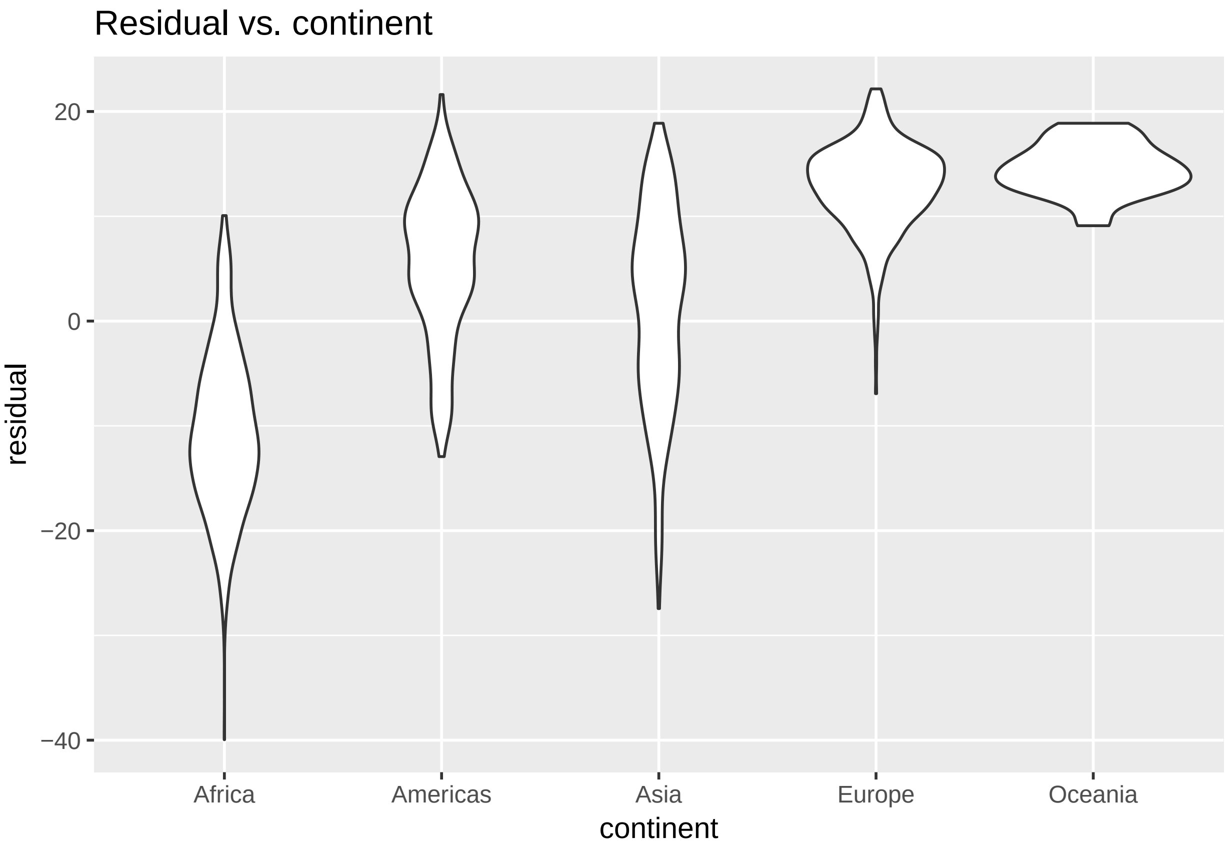
Q8: The violin plot below does match my expectations from Q4. Overall, the distribution of the plot is centered around 0 with probability distributions going from below 0 to above 0. This centering around 0 indicates a linear relationship. The plot also looks indentical to the one from Q4, except with a downward trend.

```
aug <- linear_co %>% augment()

aug %>%
  ggplot(aes(x=factor(year), y=.resid)) +
    geom_violin() + labs(title="Residual vs. year", x = "year", y = "residual")
```



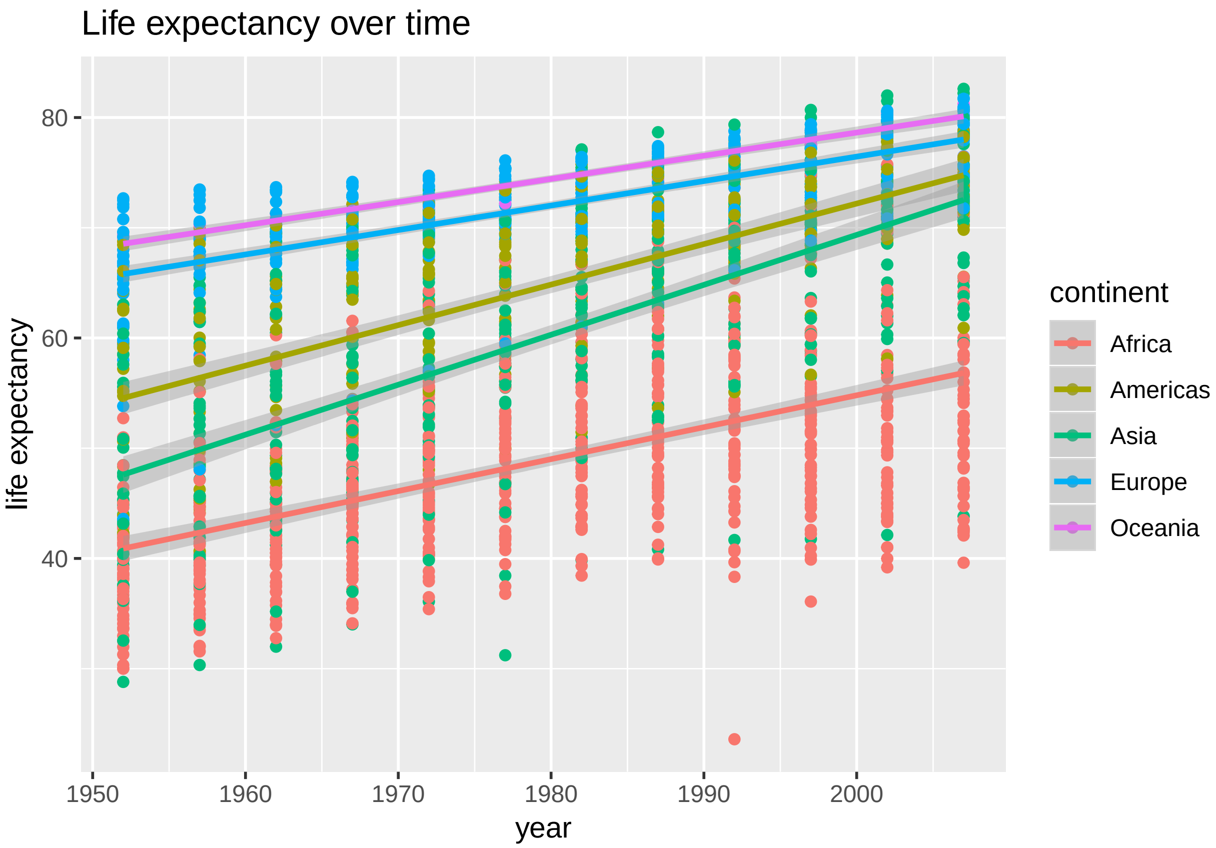
Q9: The violin plot of residual vs. continent shows that our model produces a different error for each continent. The Americas and Asia residuals closely represent the model while Africa, Europe, and Oceania does not. Africa will have a high negative error (meaning that life expectancy compared to the rest of the world is low) while Oceania and Asia have a high positive error (meaning life expectancy is high compared to the rest of the world). When performing regression analysis, the residual error of each of the continents is something to keep in mind.

```r
merged_aug <- merge(aug, gapminder, by.x="lifeExp", by.y="lifeExp")

merged_aug %>%
  ggplot(aes(x=continent, y=.resid)) +
    geom_violin() + labs(title="Residual vs. continent", x = "continent", y = "residual")
```



Q10: Based on this plot, yes there should be a term included for an interaction between continent and year. The slopes of Asia and the Americas are different than the slopes of Africa, Oceania, and europe. This indicates that life expectancy is improving at a quicker rate per year in the Americas and Asia than the three other continents. This indicates a possible interaction.

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



Q11: All parameters are signifcant from O (p.value < 0.05) except for Oceania and the year*Oceania parameters. Both have these have a p.value > 0.05, meaning that we can not reject the null hypothesis that there is no interaction. Therefore, both of these parameters are statistically insignificant.

```
interaction_co <- lm(lifeExp~year*continent, data=gapminder)
tidy_interaction_co <- interaction_co %>% tidy()
tidy_interaction_co
```

```
## # 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
```

Q12: The chart below gives the continent name and its corresponding value for how much life expectancy increases each year.

```
baseline <- tidy_interaction_co %>%filter(term == "year") %>% select(estimate)
baseline <- baseline$estimate
tidy_interaction_co <- tidy_interaction_co %>% mutate(average_per_year= ifelse((estimate != baseline),
tidy_interaction_co$term[1] = "Africa"
tidy_interaction_co$term[2] = "Americas"
tidy_interaction_co$term[3] = "Asia"
tidy_interaction_co$term[4] = "Europe"
tidy_interaction_co$term[5] = "Oceania"

tidy_interaction_co
```

```
## # A tibble: 5 x 2
##   term     average_per_year
##   <chr>               <dbl>
## 1 Africa              0.290
## 2 Americas            0.368
## 3 Asia                0.453
## 4 Europe              0.222
## 5 Oceania             0.210
```

Q13: Overall, both models have low enough p-values ($p < 0.05$) where they are statistically significant. However, the residuals in the interaction model have a MSE of 52, compared with the MSE of 13 in the original linear regression model (as seen in the two charts below). The mean square of the error is calculated by dividing the sum of squares of the residual error by the degrees of freedom.

F-Testing the linear regression model from Q6

```
linear_co <- lm(lifeExp~year, data=gapminder)
anova(linear_co)
```

```
## 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
```

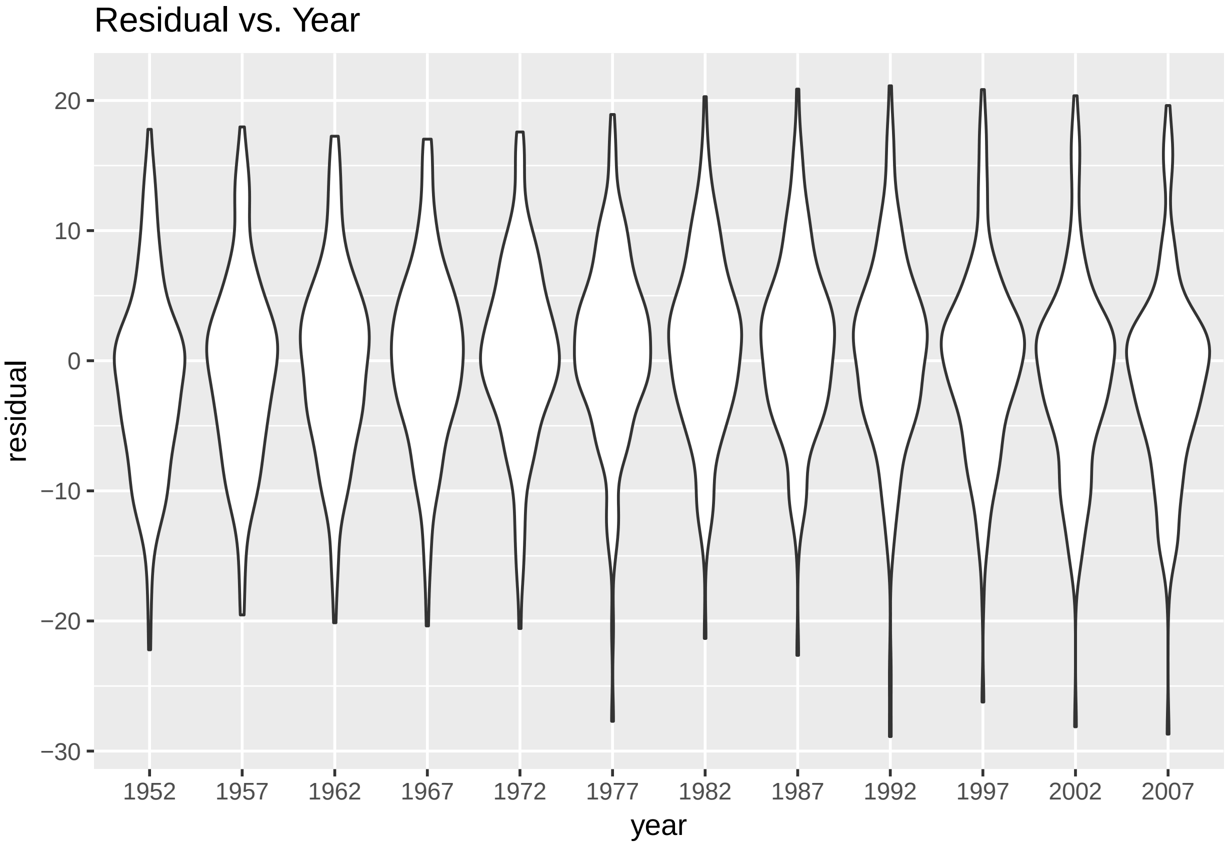
F-Testing the interaction model from Q11

```
interaction_co <- lm(lifeExp~year*continent, data=gapminder)
anova(interaction_co)
```

```
## 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
```

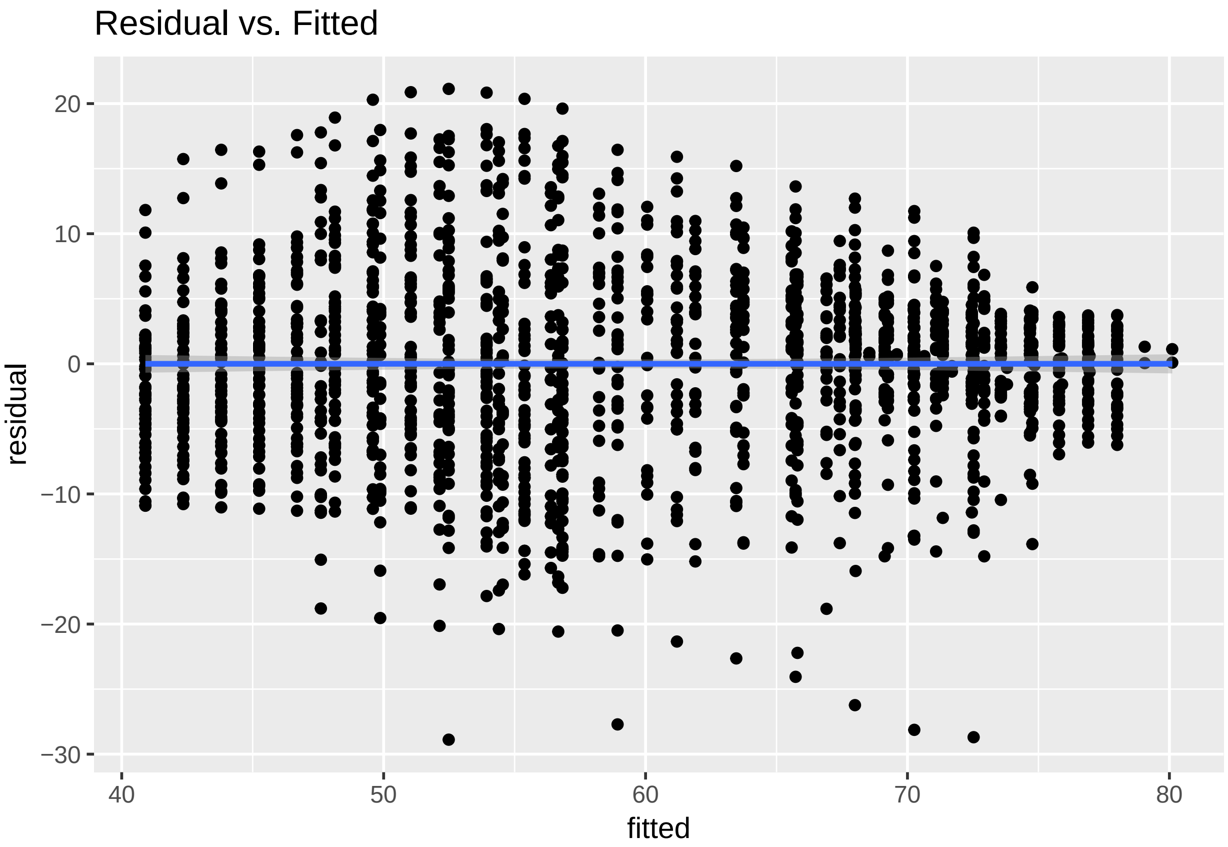
To further prove the statement above, a residuals vs. year violin plot shows how well the interaction model matches assumptions of a linear regression model. Each year is represented by a unimodal violin plot with the probability being most densely distribution around a residual value of 0. This means that, overall, the average difference between the actual and predicted value is small (and much smaller than the original linear regression model).

```
interaction_co %>% augment() %>% ggplot(aes(x=factor(year), y=.resid)) +
    geom_violin() + labs(title="Residual vs. Year", x = "year", y = "residual")
```



A scatter plot of residuals vs. fitted values shows that the underlying relationship is linear. By using a linear regression line, we can see that the residual values are centered around 0. Again, this indicates that the average difference between the actual and predicted value is centered around 0.

```
interaction_co %>% ggplot(mapping=aes(x=.fitted, y=.resid)) + geom_point() + geom_smooth(method=lm) + la
```


Residual vs. Fitted
residual
20
10
0
−10
−20
−30
40
50
60
70
80
fitted