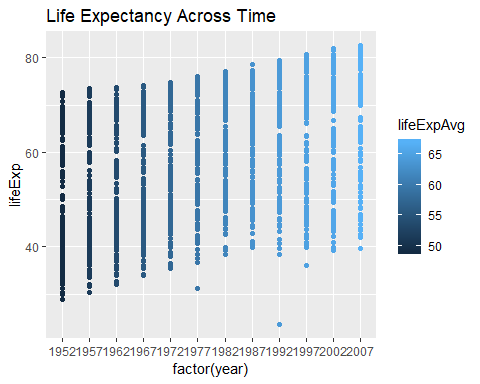
Project3

Michael Donahue

May 2, 2019

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

g <- gapminder  
  
g <- group\_by(g, year) %>% mutate(lifeExpAvg = mean(lifeExp))  
  
g %>%   
 ggplot() +  
 geom\_point(aes(x=factor(year), y=lifeExp, color=lifeExpAvg)) +   
 labs(title="Life Expectancy Across Time")



# 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)

There is an increase in life expectancy over time. As shown by this graph, the cnetral tendency increases from the high 40s in 1952 to slightly over 65 in 2007. This increase “”

# 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?

Taking the central tendency as the median to check if the data is skewed, it is pretty obviously skewed over all years with only 1972 being close.In the first half of the graph, it is skewed below the median, in 1972, it changes and it skewed in the upper half.

# 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, you reject the relationship of no relationship. I believe there is a relationship since, as years increase, you see a slight increase in the averages of life exp.

# 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)

I believe you will see a violin plot where each year has a similar shape to the one given. However, there will be downward trend as the plots move toward 2007 since the data is closer together in later years.

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

It will hold the same shape, as well it will tend toward zero on the side with more data clustered together. If the data is not skewed, you’ll see a semetraical shape with zero as the cut.

# 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.

linear\_model <- lm(lifeExp~year, data=g)  
broom::tidy(linear\_model)

## # 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

# Q6: On average, by how much does life expectancy increase every year around the world?

It increases .3259 a year

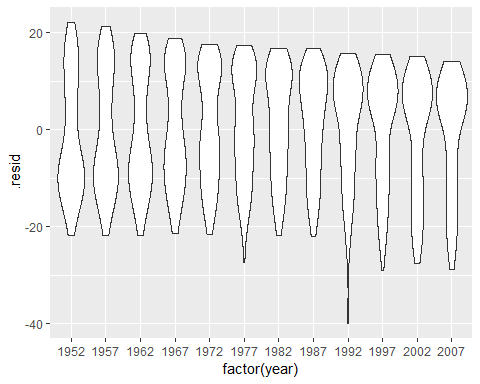
# Q7: Do you reject the null hypothesis of no relationship between year and life expectancy? Why?

Yes, we reject the Null Hypothesis.

Since B1 != 0 (it equals .3259), there is a relationship between years and lifeExp. As years increase, so does the life Exp.

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

linear\_model <- lm(lifeExp~year, data=g)  
  
broom::augment(linear\_model) %>%  
 ggplot(aes(x=factor(year), y=.resid)) +  
 geom\_violin()

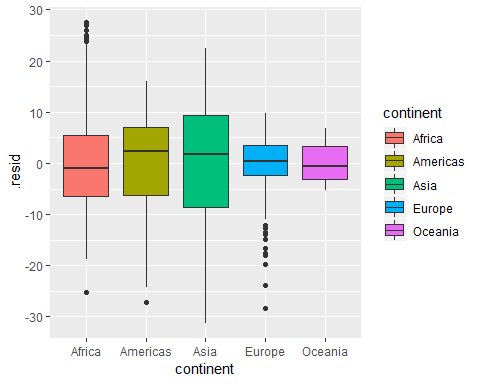


# Question 8: Does the plot of Exercise 3 match your expectations (as you answered Question 4)?

It seems to have matched my overall guess. I was thinking the latter plots would be closer to zero. However, it has the correct shape and correct trajectory.

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

linear\_model <- lm(lifeExp~continent, data=g)  
  
broom::augment(linear\_model) %>%  
 ggplot(aes(x=continent, y=.resid, fill=continent)) +  
 geom\_boxplot()

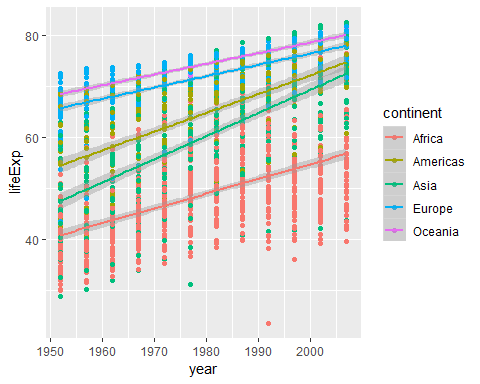


# Question 9: Is there a dependence between model residual and continent? If so, what would that suggest when performing a regression analysis of life expectancy across time?

There is a dependency between residuals and continent. Since life exp increases over time, you could argue a migration possibly happened to countries where life exp is higher.

# 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).

g %>% group\_by(continent) %>%  
 ggplot(aes(x=year, y=lifeExp, color=continent)) +  
 geom\_point() +  
 geom\_smooth(method=lm)



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

I believe it should since Life Exp is dependent on both year and country.

# 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.

linear\_model <- lm(lifeExp~year\*continent, data=g)  
broom::tidy(linear\_model)

## # 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?

I believe all of the continents’ slopes are signigicantly different than zero with Africa and Asia the most signigicant.

# 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)

broom::tidy(linear\_model) %>% slice(6:10) %>% select(term, estimate, p.value)

## # A tibble: 5 x 3  
## term estimate p.value  
## <chr> <dbl> <dbl>  
## 1 continentOceania 182. 0.287   
## 2 year:continentAmericas 0.0781 0.00758   
## 3 year:continentAsia 0.164 0.00000000115  
## 4 year:continentEurope -0.0676 0.0142   
## 5 year:continentOceania -0.0793 0.360

# Exercise 7: Use the anova function to perform an F-test that compares how well two models fit your data: (a) the linear regression models from Exercise 2 (only including year as a covariate) and (b) Exercise 6 (including interaction between year and continent).

other\_model <- lm(lifeExp~year, data=g)  
fit <- anova(other\_model, linear\_model)   
  
fit

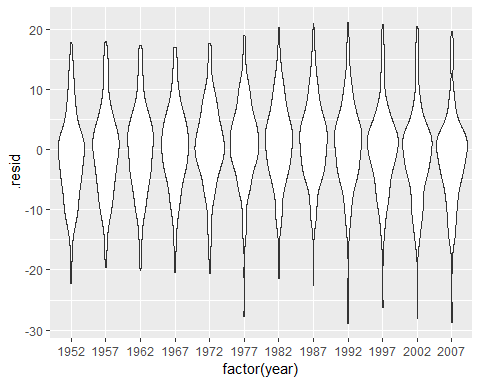
## Analysis of Variance Table  
##   
## Model 1: lifeExp ~ year  
## Model 2: lifeExp ~ year \* continent  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 1702 230229   
## 2 1694 87320 8 142909 346.55 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Question 13: Is the interaction model significantly better than the year-only model? Why?

The first model (the one that includes with only year) has a much higher residual. Thus, the second model has a much better relationship with life expectancy.

# Exercise 8: 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. (You should use the broom::augment function).

broom::augment(linear\_model) %>%  
 ggplot(aes(x=factor(year), y=.resid)) +  
 geom\_violin()



This seems fairly accurate since the increase in accuracy will cause a more evenly distributed distance to the linear model. As well, each year is clustered closer to zero since the amount of error is less. This proves using the year \* continent relationship is the better model.