

## Factors Influencing Life Expectancy

Lauren Flemmer

STAT 230A

### *I. Introduction*

Life expectancy is an important measure of the overall health of a population, and is used widely to inform many of the decisions policy makers and government authorities make regarding that population. Since many factors could possibly have an effect on the life expectancy of a population, I found the problem of modeling these relationships interesting. The dataset I chose to use for this project was collected from the World Health Organization, and relates to the life expectancy of 193 countries from the years 2000-2015. For each country in each year, there is data for factors such as alcohol consumption, infant deaths, health expenditure, measles cases, BMI, prevalence of thinness, and years of schooling, among others. Since life expectancy is a strong indicator of a country's economic development and overall health, being able to accurately model relationships between various factors and life expectancy has the potential to inform and drive changes within these countries by governments and policymakers. By understanding how various factors might contribute to a country's life expectancy, governments and institutions would be able to build infrastructure, create policies, and better allocate resources to countries that have low life expectancy, and hopefully address the factors that are negatively contributing to that issue. Upon choosing this topic and dataset, the two questions I was interested in answering were the following. First, which personal/cultural habits, such as alcohol consumption, are most detrimental to a country's life expectancy? Second, what relationships do immunizations against health conditions such as measles, polio, and HIV have with life expectancy? By conducting this analysis, I was able to thoroughly answer both of these questions, and gain a strong understanding of the factors that influence the life expectancy of a country.

### *II. Modeling*

Because the structure of the data includes measures for each country over a number of years, there is clearly some sort of autocorrelative structure among observations from each country. In other words, the life expectancy of a country in a given year is likely dependent on and correlated with that of the previous years' observations. Because this dataset has a clustered, longitudinal structure, I utilized the "geepack" package in R, implementing methods proposed by Liang and Zeger in their 1986 paper, where they introduced the idea of a Generalized Estimating Equation for fitting GLMs to data with this same structure. Before modeling this life expectancy data, to determine which features belonged in the initial model, I created a correlation matrix and plotted features against life expectancy. The features that were moderately and

strongly correlated with life expectancy were included in the initial model. After fitting this initial model, I performed backwards selection, removing all features that were not statistically significant in predicting life expectancy. After backwards selection, the features that remained statistically significant in the model were the development status (“developed” or “developing”), the year, HIV/AIDS deaths per 1000 live births, the Human Development Index (in terms of income composition of resources), and the number of years of schooling. An important feature of this method of modeling longitudinal data is the working covariance matrix of the model, which represents the relationship among observations within each country. In order to build the final model in this analysis, I had to experiment with different correlation structures, in order to find the method that produced the best results, and made the most sense for the problem at hand. After testing the “independent”, “unstructured”, “exchangeable”, and “AR-1” correlation structures on the final model reduced by backwards selection, I chose to use the AR-1 structure to represent the within-cluster correlation. AR-1, which assumes that measurements within a cluster have an autoregressive relationship, made the most sense when considering the life expectancy data, since the observations within each country are assumed to have an autocorrelative relationship. Additionally, the estimates in the AR-1 model had the smallest standard errors and negligible mean absolute error when compared to the other correlation structures. The results of this final model will be discussed in the next section.

### *III. Discussion of Results*

As discussed in the last section, the final model chosen to model human life expectancy was a GLM with a Gaussian link function, using the AR-1 correlation structure. In this final model, all predictors were statistically significant, so by looking at the coefficient estimates of the model, we are able to understand the relative importance of the features in the model, and identify how much each feature contributes to life expectancy. Of all the predictors in the final model, the development status of the country had the largest coefficient estimate,  $b_1 = -5.9522$ , meaning that a country development static changing from “developed” to “developing”, all other predictors held constant, will decrease the life expectancy of that country by approximately 6 years. This value is significantly larger than that of the other predictors, which allows us to understand the importance of this feature, and impact it has on life expectancy compared to other features. The second most influential predictor in this model was the number of years of schooling in the country, with  $b_2 = 1.1141$ . Therefore, for each additional year of schooling in a particular country, we can expect its life expectancy to increase by 1.1141, all other factors held constant. The next most influential predictors in the model were the HIV/AIDS deaths per 1000 live births, the year, and the Human Development Index (in terms of income composition of resources), respectively. These coefficient estimates all had absolute values smaller than 1, so their relative influences on life expectancy are not

as large as the influences of the development status and years of schooling. The exact coefficient values in this model can be found in the Appendix. The standard errors of these estimates were generally smaller than the standard errors for the same model, with an independent (instead of AR-1) correlation structure. This illustrates the utility of the AR-1 correlation structure in this particular application, since it produces more precise estimates for the model coefficients.

#### *IV. Conclusion*

Looking back at the two questions posed at the beginning of this project, the analysis and modeling conducted allow us to answer both of these questions. First, I posed the question: Which personal or cultural habits are most detrimental to a country's life expectancy? Looking at the results from the final model, we can observe that no personal or cultural habits, such as alcohol consumption, are present as predictors. From this, we can conclude that the life expectancy of a country is less dependent on the behaviors of the population than it is on the larger-scale infrastructure and development of the country. The second question posed before completing this project was: What are the relationships between immunization against health conditions and life expectancy? The life expectancy data used includes features that represent the percentage of the population who have immunization for diseases such as Polio, Hepatitis B, Measles, and Diphtheria, as well as a variable representing the number of deaths (ages 0-4) from HIV/AIDS. If these features were significant in the final model, we would be able to conclude that immunizations, or a lack thereof, have an impact on the overall life expectancy of a country. When looking at the results from the final model, we can see that of all these health-related predictors, HIV/AIDS deaths is the only significant predictor, implying that a lack of immunization against HIV negatively affects the life expectancy of a country. Overall, no personal or cultural habits were particularly significant or important in predicting the life expectancy of a country. As most of the significant predictors are related to the infrastructure and development of countries, we can conclude that a particular country's life expectancy is predominantly determined by education, immunization, and resource utilization. Thus, in order to address inequities in human life expectancy, governments and policy makers should focus on improving the educational and administrative structure of these countries, by improving access to and length of educational programs, and restructuring institutions and parties that manage and allocate resources. Additionally, dedicating more effort and capital toward immunization programs would undoubtedly improve the life expectancy of that country.

#### *V. Additional Analysis*

Although the data was already clustered in a way that made a lot of sense, I considered the fact that some countries are very similar in terms of overall development, and thought about trying out a new clustering scheme for the data. Rather than clustering

by country, I experimented with clustering the most similar observations together in what I hoped would become “groups of similar countries” as the new clusters. In order to create these new “similar country” clusters, I used K-Means clustering to cluster individual observations together based on their “similarity” in the Euclidean space. In order to choose the optimal number of clusters, I tuned the parameter K, choosing K such that the total within-cluster sum of squares was minimized. The optimal number of clusters ended up being K=10, with each cluster containing “similar” observations. The list of unique countries belonging to each of these clusters can be found in the Appendix. My initial expectation was for these clusters to represent groups of countries from similar regions, with similar levels of development and infrastructure. Although some of the clusters make sense, many countries have observations from different years that are split up among two or more clusters. Therefore, these clusters make more sense being interpreted as containing the most similar observations at the yearly level, rather than the country level, since individual observations from countries represent yearly measures. In order to determine if these new clusters were more informative than the original clustering by country, I fit a new model, now specifying the new K-Means clusters as the model’s clusters. Because these new clusters do not have any longitudinal structure, i.e, the observations within each cluster are not repeated measurements over a period of time, I didn’t expect for them to have any autocorrelative structure. Thus, I used the independent correlation structure for this model. After fitting this model using the new clusters, I compared it to my initial/full (prior to any backward selection) AR-1 model which was clustered by country. Looking at the regression output, the AR-1 country-clustered model had coefficient estimate P-values that were significantly smaller than the P-values of the K-means clustered model. Additionally, all but one of the standard errors in the country-clustered model were significantly smaller than that of the K-means clustered model. Therefore, the original choice to cluster by country in the model seems to be a much more robust choice for modeling life expectancy. Although this new method of clustering didn’t perform very well, it was interesting to try out. Since these new clusters appear relatively random, they don’t seem to be adding much to the clustered model- therefore the model fit using these clusters might be very similar to a regular GLM.

*VI. References*

1. KUNG-YEE LIANG, SCOTT L. ZEGER, Longitudinal data analysis using generalized linear models, *Biometrika*, Volume 73, Issue 1, April 1986, Pages 13–22, <https://doi.org/10.1093/biomet/73.1.13>
2. “Generalized Estimating Equations.” *IBM*, [www.ibm.com/docs/en/spss-statistics/saas?topic=statistics-generalized-estimating-equations](http://www.ibm.com/docs/en/spss-statistics/saas?topic=statistics-generalized-estimating-equations). Accessed 6 May 2023.
3. “12.1 - Introduction to Generalized Estimating Equations: Stat 504.” *PennState: Statistics Online Courses*, [online.stat.psu.edu/stat504/lesson/12/12.1](http://online.stat.psu.edu/stat504/lesson/12/12.1). Accessed 6 May 2023.
4. “Geepack: Generalized Estimating Equation Package.” *CRAN*, [cran.r-project.org/web/packages/geepack/geepack.pdf](http://cran.r-project.org/web/packages/geepack/geepack.pdf). Accessed 9 May 2023.

# Code Appendix

## Read in data

```
lifeExpectancy <- read.csv('/Users/laurenflemmer/Desktop/life_expectancy_proj/Life Expectancy Data.csv')

lifeExpectancy <- lifeExpectancy %>% drop_na()

head(lifeExpectancy, 5)

##      Country Year Status Life.expectancy Adult.Mortality infant.deaths
## 1 Afghanistan 2015 Developing          65.0           263            62
## 2 Afghanistan 2014 Developing          59.9           271            64
## 3 Afghanistan 2013 Developing          59.9           268            66
## 4 Afghanistan 2012 Developing          59.5           272            69
## 5 Afghanistan 2011 Developing          59.2           275            71
##   Alcohol percentage.expenditure Hepatitis.B Measles   BMI under.five.deaths
## 1     0.01                71.279624       65    1154 19.1             83
## 2     0.01                73.523582       62    492 18.6             86
## 3     0.01                73.219243       64    430 18.1             89
## 4     0.01                78.184215       67   2787 17.6             93
## 5     0.01                7.097109       68   3013 17.2             97
##   Polio Total.expenditure Diphtheria HIV.AIDS        GDP Population
## 1     6               8.16          65  0.1 584.25921  33736494
## 2    58               8.18          62  0.1 612.69651  327582
## 3    62               8.13          64  0.1 631.74498  31731688
## 4    67               8.52          67  0.1 669.95900  3696958
## 5    68               7.87          68  0.1 63.53723  2978599
##   thinness..1.19.years thinness.5.9.years Income.composition.of.resources
## 1                 17.2          17.3                      0.479
## 2                 17.5          17.5                      0.476
## 3                 17.7          17.7                      0.470
## 4                 17.9          18.0                      0.463
## 5                 18.2          18.2                      0.454
##   Schooling
## 1      10.1
## 2      10.0
## 3      9.9
## 4      9.8
## 5      9.5

colSums(is.na(lifeExpectancy))

##                  Country                   Year
##                  0                         0
##                  Status                 Life.expectancy
##                  0                         0
##          Adult.Mortality         infant.deaths
##                  0                         0
```

```

##          Alcohol           percentage.expenditure
##          0                   0
##          Hepatitis.B        Measles
##          0                   0
##          BMI                under.five.deaths
##          0                   0
##          Polio              Total.expenditure
##          0                   0
##          Diphtheria         HIV.AIDS
##          0                   0
##          GDP                Population
##          0                   0
##          thinness..1.19.years thinness.5.9.years
##          0                   0
## Income.composition.of.resources Schooling
##          0                   0

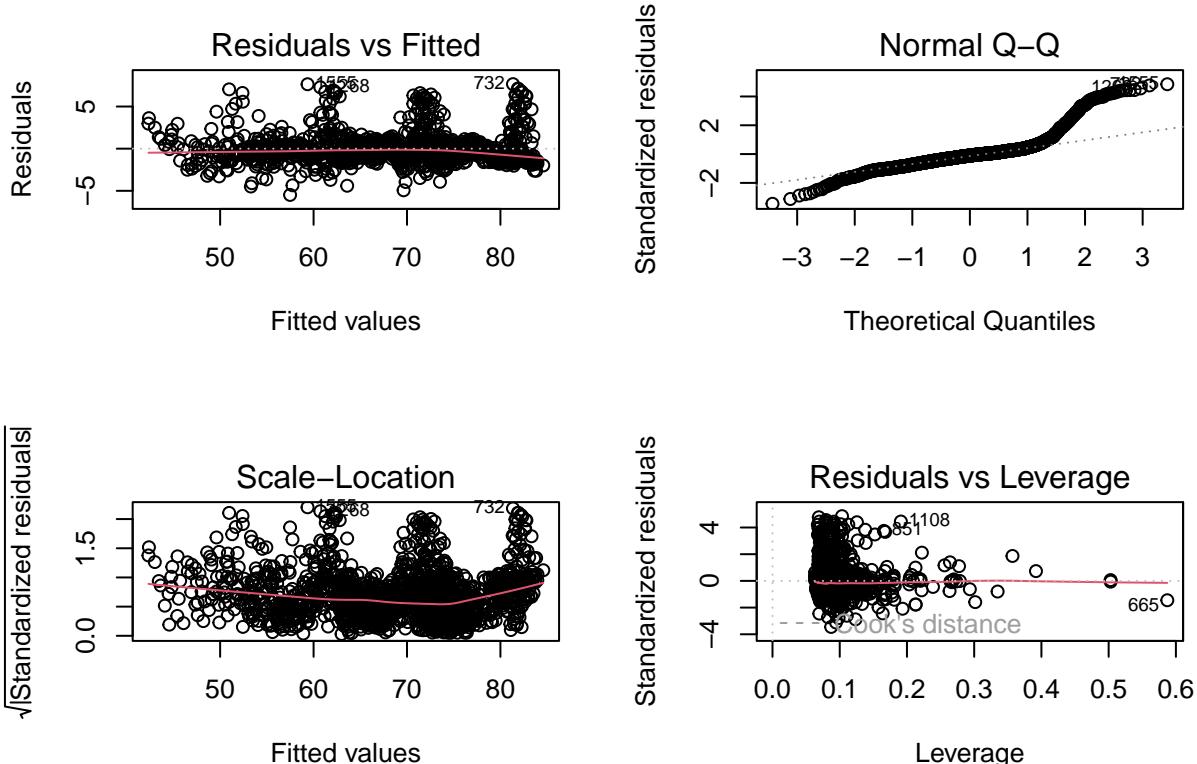
```

To determine which model is appropriate, plot diagnostics

```

par(mfrow=c(2,2))
# diagnostic plots
basic_model <- lm(Life.expectancy ~ ., data = lifeExpectancy)
plot(basic_model)

```

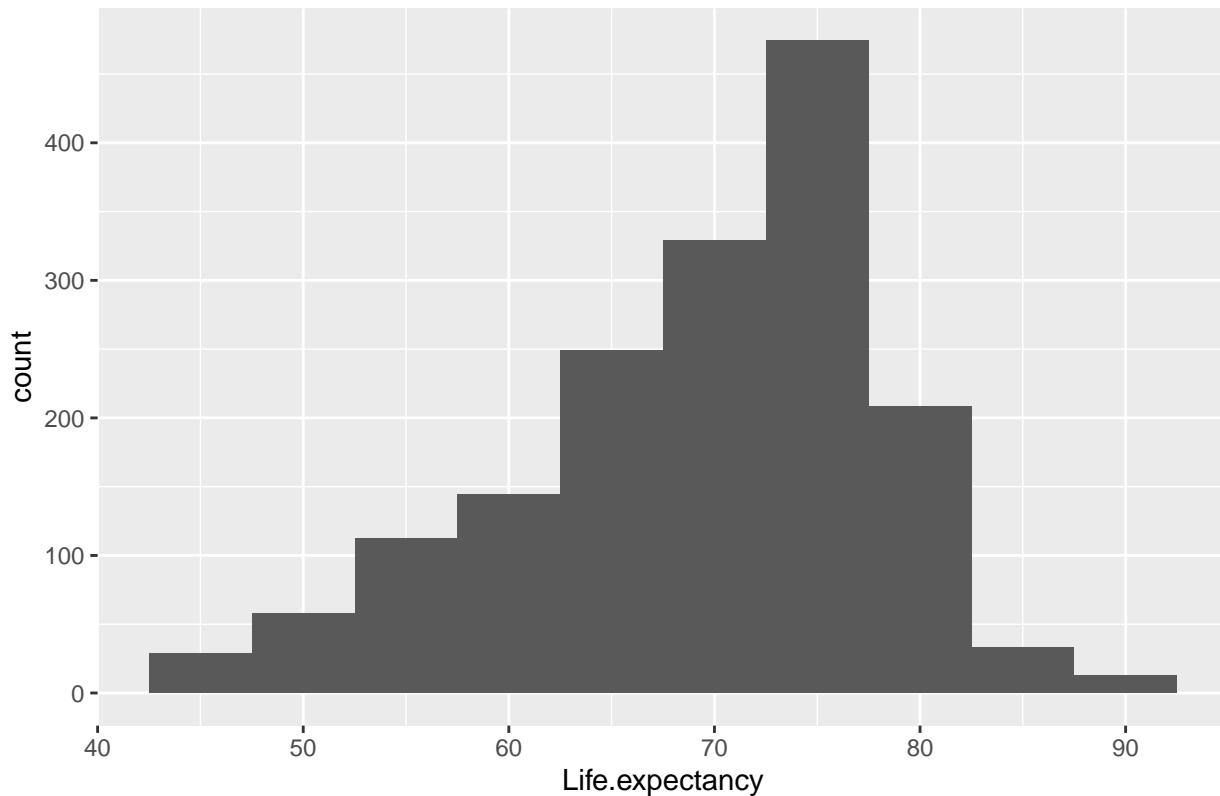


```

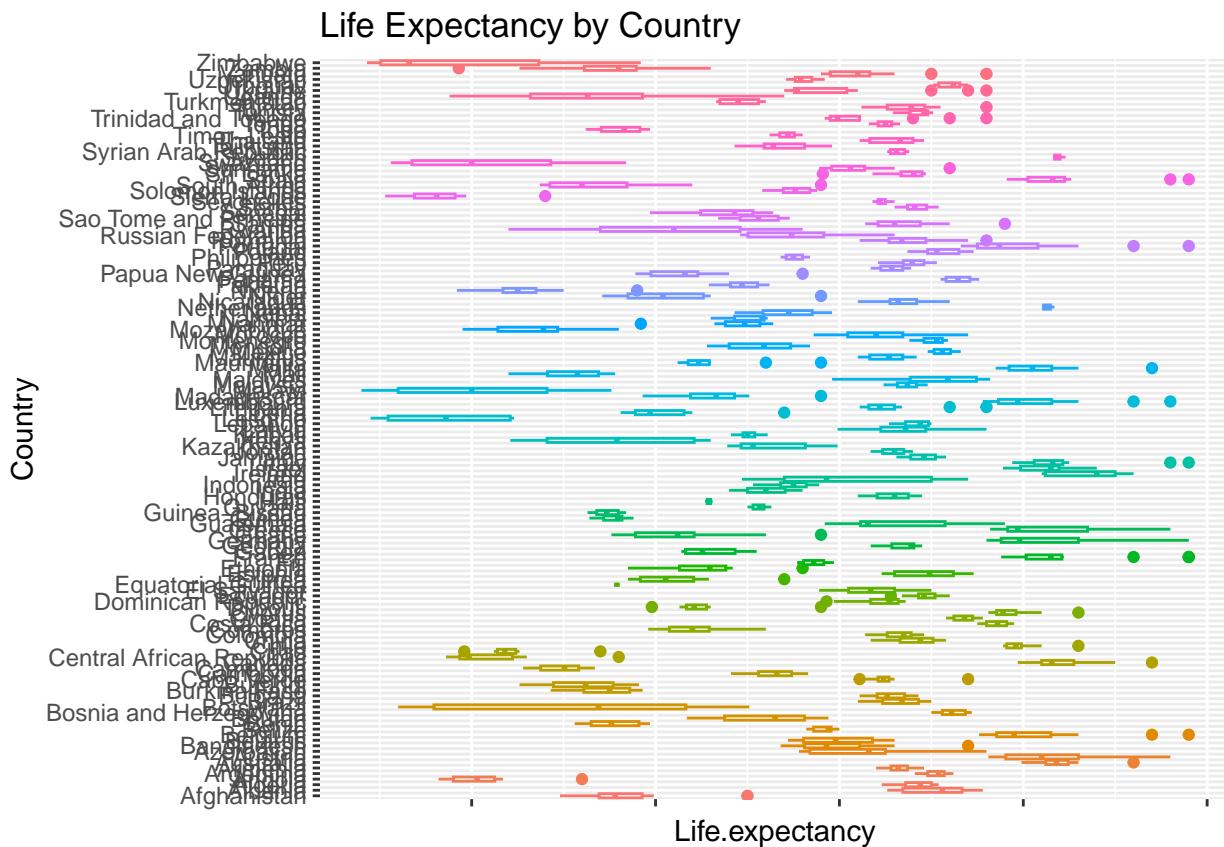
# plot response
ggplot(data=lifeExpectancy) +
  geom_histogram(mapping=aes(x=Life.expectancy), bins=10) +
  ggtitle("Distribution of life expectancy")

```

## Distribution of life expectancy

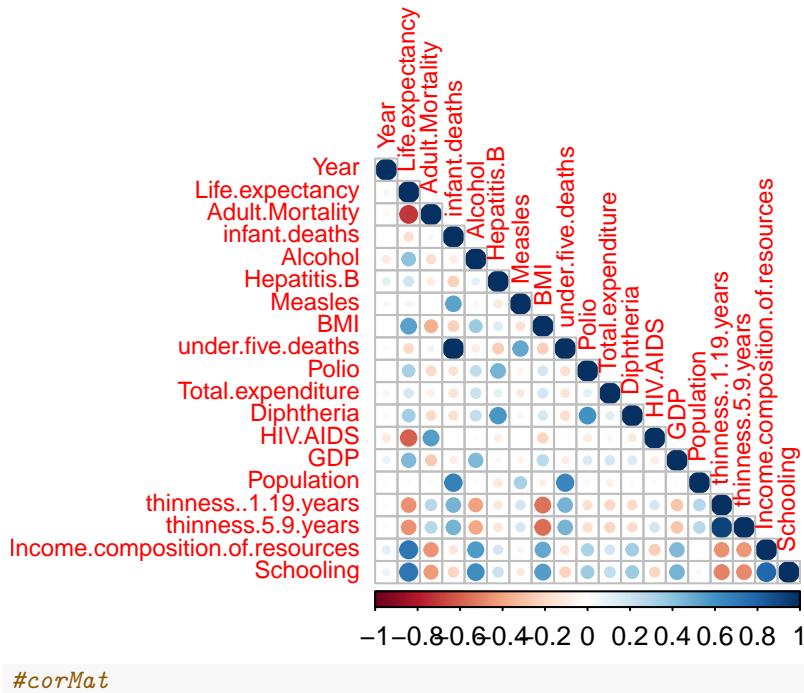


```
ggplot(data=lifeExpectancy) +  
  geom_boxplot(mapping=aes(x=Country, y=Life.expectancy, color=Country), bins=10) +  
  ggtitle("Life Expectancy by Country") +  
  theme(legend.position = "none") +  
  coord_flip() +  
  theme(axis.text.x = element_text(angle = 80, vjust = 0.5, hjust=1, size=0.01))
```



```
# correlation plot
# include only continuous vars
# removed percentage expenditure bc redundant
continuous_vars <- lifeExpectancy %>% select(Year, Life.expectancy,
                                               Adult.Mortality, infant.deaths,
                                               Alcohol, Hepatitis.B, Measles, BMI,
                                               under.five.deaths, Polio, Total.expenditure,
                                               Diphtheria, HIV.AIDS, GDP, Population,
                                               thinness..1.19.years, thinness..5.9.years,
                                               Income.composition.of.resources, Schooling)

corMat <- cor(continuous_vars)
corrplot(corMat, type="lower", method="circle", tl.cex = 0.75)
```



```
#corMat
```

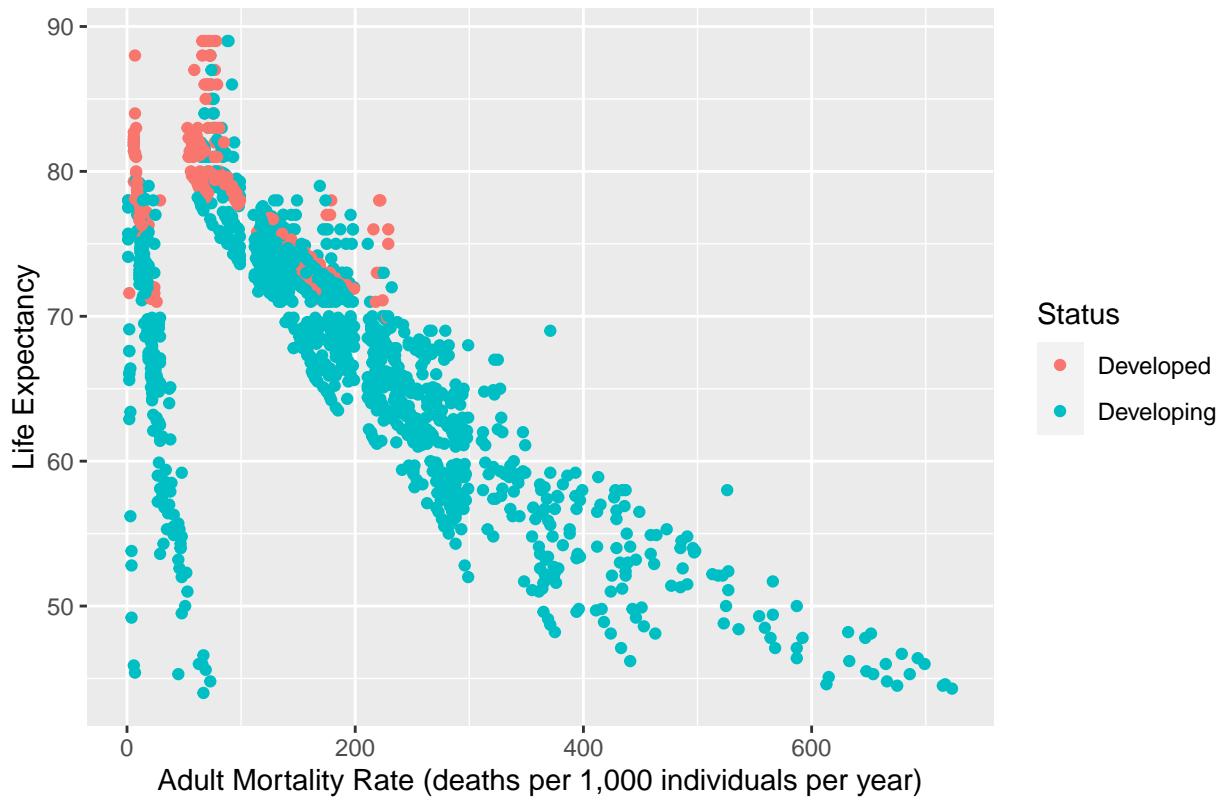
Variables correlated with life expectancy ( $|r| > 0.4$ ):

- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- GDP
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

## EDA for variables correlated with life expectancy

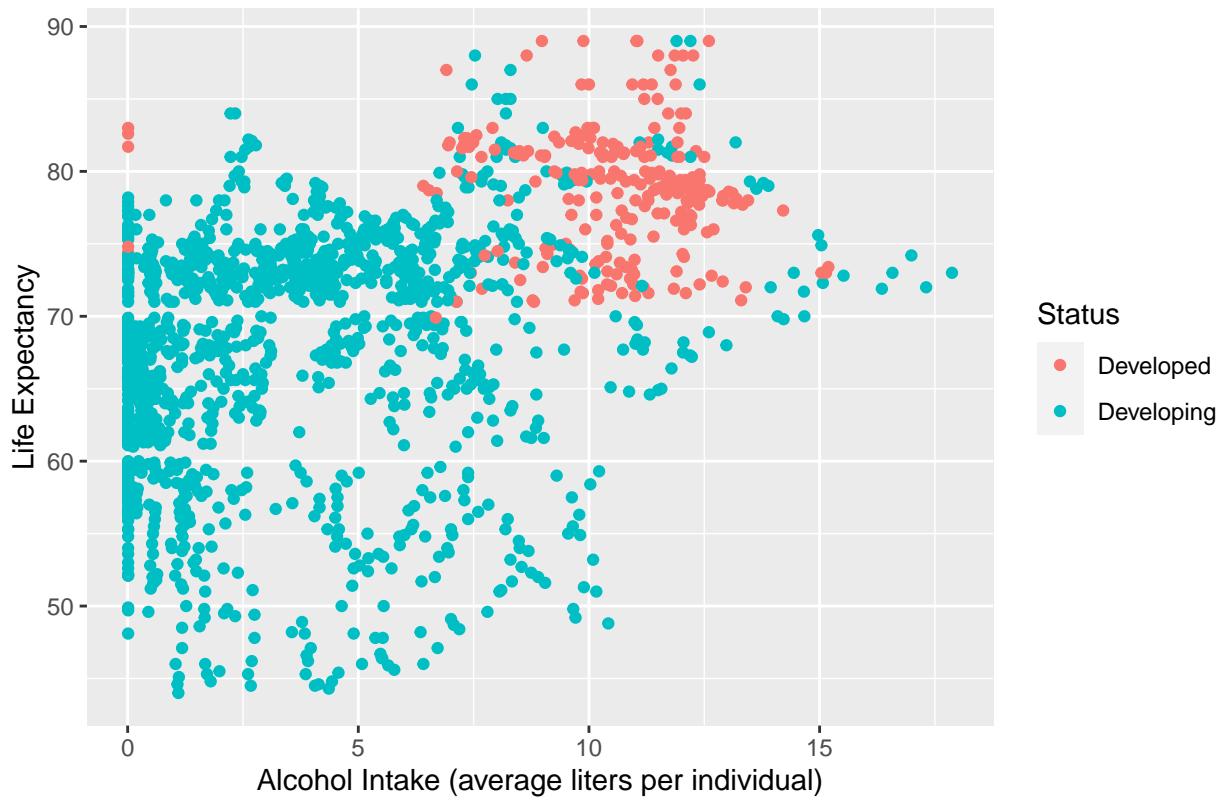
```
# adult mortality vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Adult.Mortality, y=Life.expectancy, color>Status)) +
  ggtitle("Adult Mortality Rate vs. Life Expectancy") +
  xlab("Adult Mortality Rate (deaths per 1,000 individuals per year)") +
  ylab("Life Expectancy")
```

## Adult Mortality Rate vs. Life Expectancy

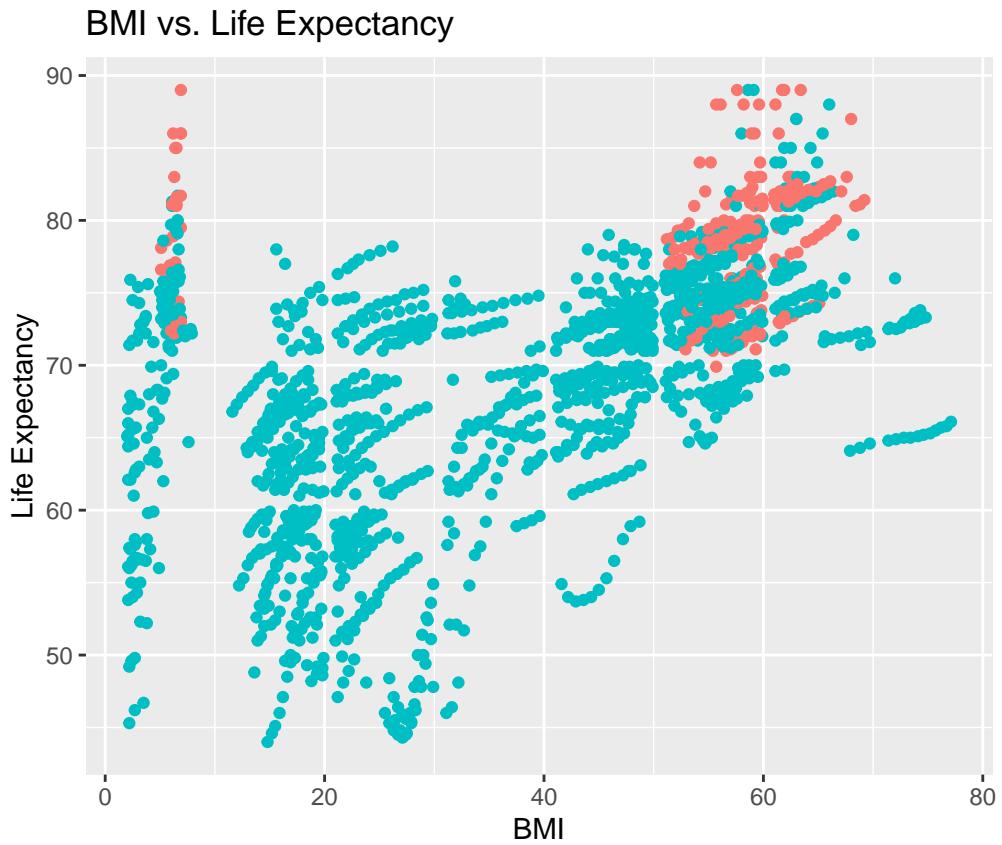


```
# alcohol intake vs life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Alcohol, y=Life.expectancy, color>Status)) +
  ggtitle("Alcohol Intake vs. Life Expectancy") +
  xlab("Alcohol Intake (average liters per individual)") +
  ylab("Life Expectancy")
```

## Alcohol Intake vs. Life Expectancy

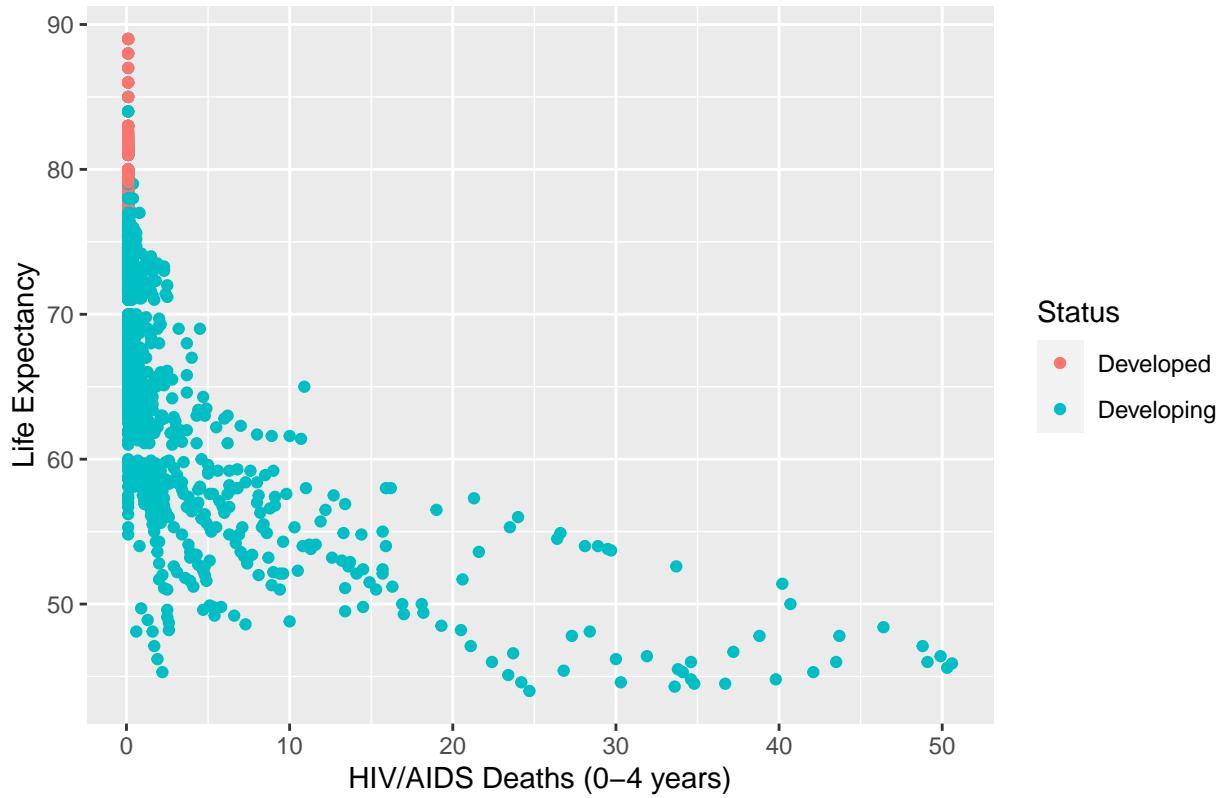


```
# bmi vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=BMI, y=Life.expectancy, color>Status)) +
  ggtitle("BMI vs. Life Expectancy") +
  xlab("BMI") +
  ylab("Life Expectancy")
```



```
# hiv/aids vs life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=HIV.AIDS, y=Life.expectancy, color=Status)) +
  ggtitle("HIV/AIDS Deaths vs. Life Expectancy") +
  xlab("HIV/AIDS Deaths (0-4 years)") +
  ylab("Life Expectancy")
```

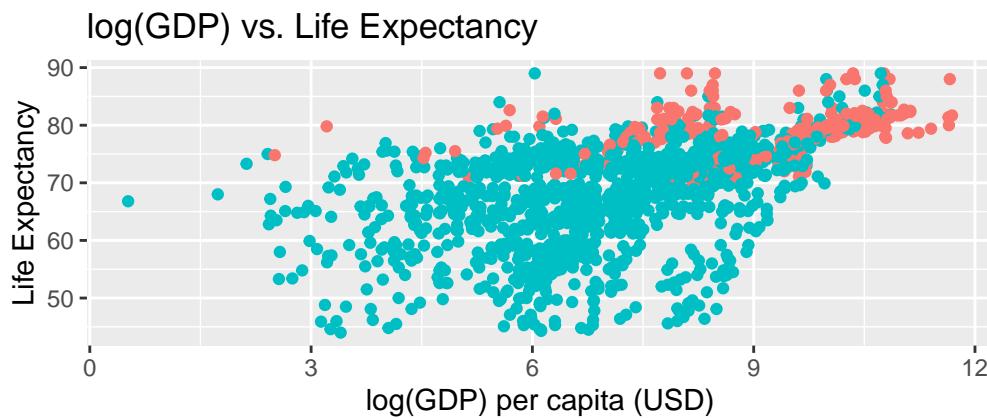
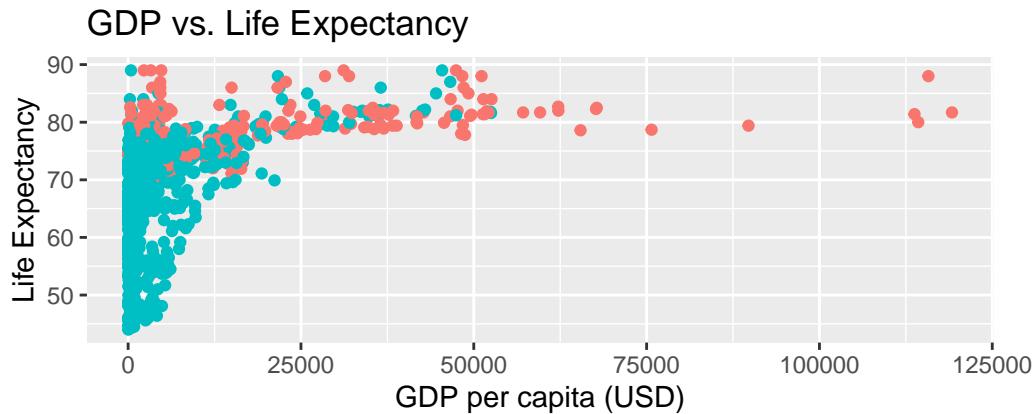
## HIV/AIDS deaths vs. Life Expectancy



```
# gdp vs life expectancy
gdp_plot <- ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=GDP, y=Life.expectancy, color>Status)) +
  ggtitle("GDP vs. Life Expectancy") +
  xlab("GDP per capita (USD)") +
  ylab("Life Expectancy")

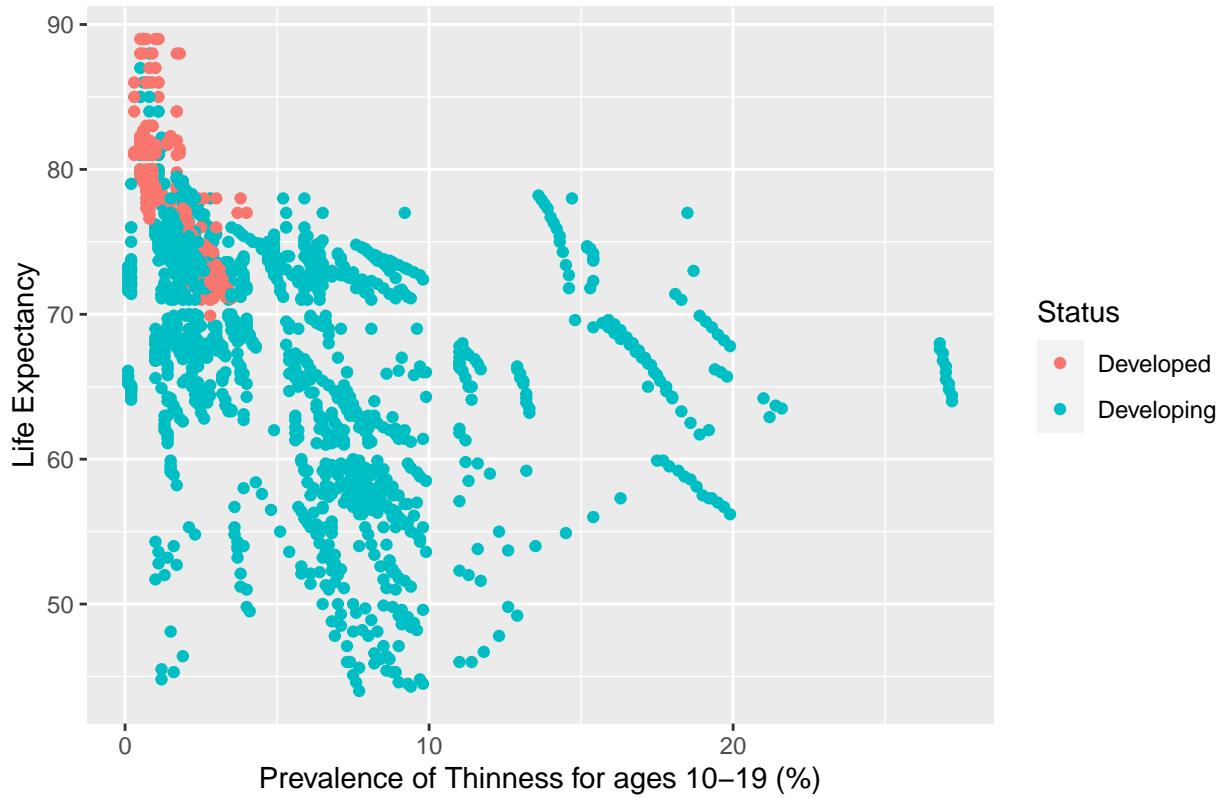
gdp_log_plot <- ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=log(GDP), y=Life.expectancy, color>Status)) +
  ggtitle("log(GDP) vs. Life Expectancy") +
  xlab("log(GDP) per capita (USD)") +
  ylab("Life Expectancy")

grid.arrange(gdp_plot, gdp_log_plot)
```



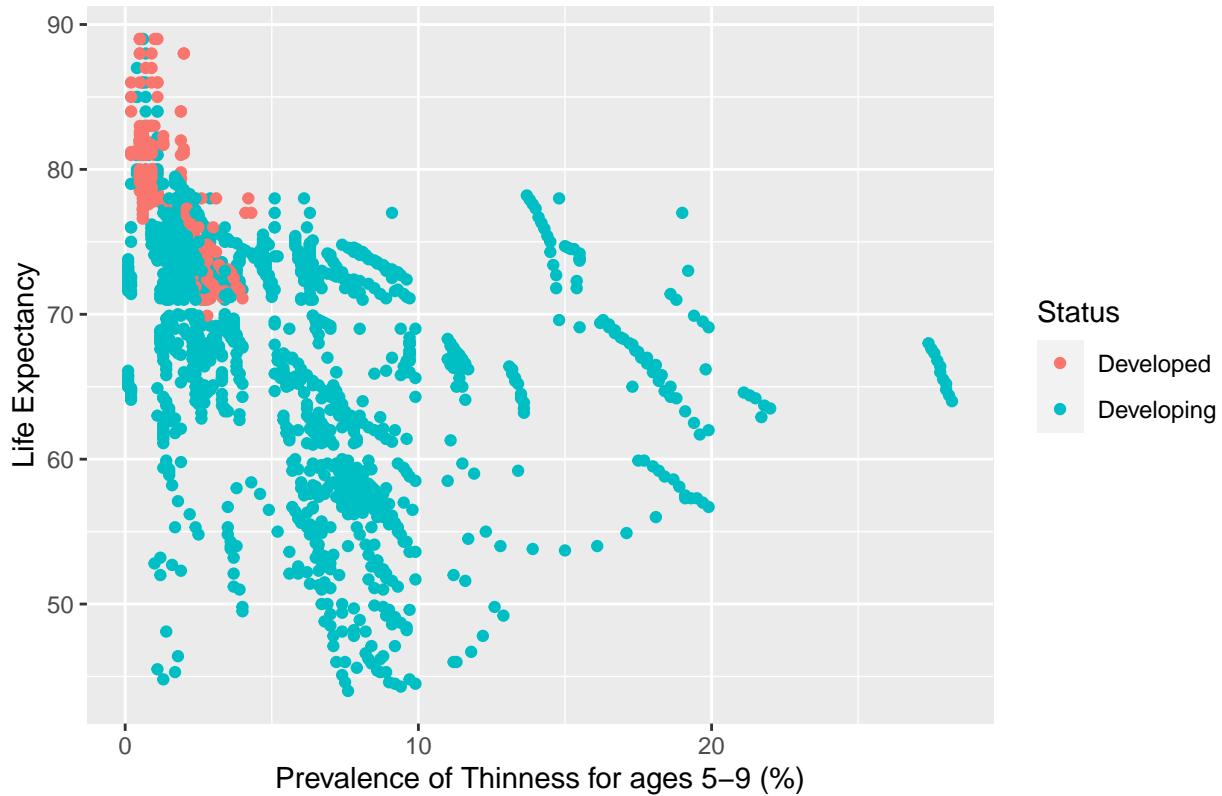
```
# prevalence of thinness (ages 10-19)
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=thinness..1.19.years, y=Life.expectancy, color>Status)) +
  ggtitle("Prevalence of Thinness (ages 10-19) vs. Life Expectancy") +
  xlab("Prevalence of Thinness for ages 10-19 (%)") +
  ylab("Life Expectancy")
```

## Prevalence of Thinness (ages 10–19) vs. Life Expectancy



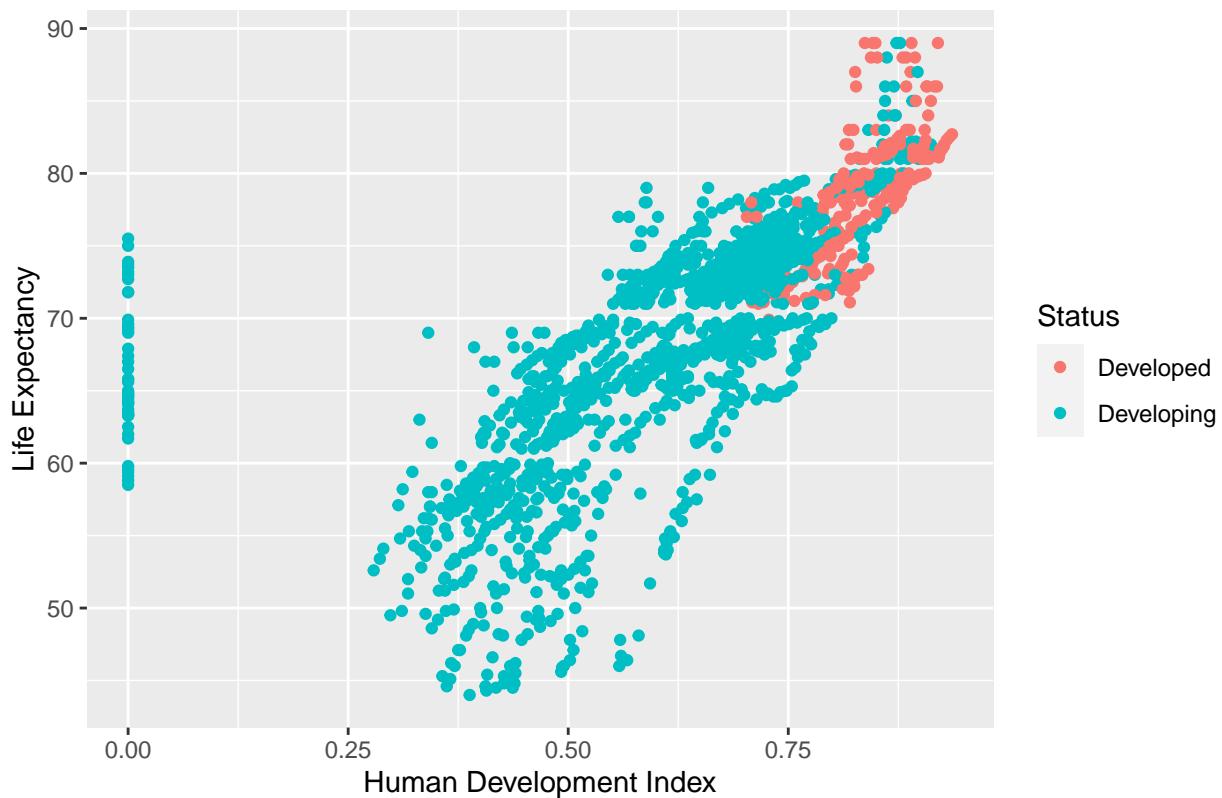
```
# prevalence of thinness (ages 5-9)
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=thinness.5.9.years, y=Life.expectancy, color>Status)) +
  ggtitle("Prevalence of Thinness (ages 5-9) vs. Life Expectancy") +
  xlab("Prevalence of Thinness for ages 5-9 (%)") +
  ylab("Life Expectancy")
```

## Prevalence of Thinness (ages 5–9) vs. Life Expectancy



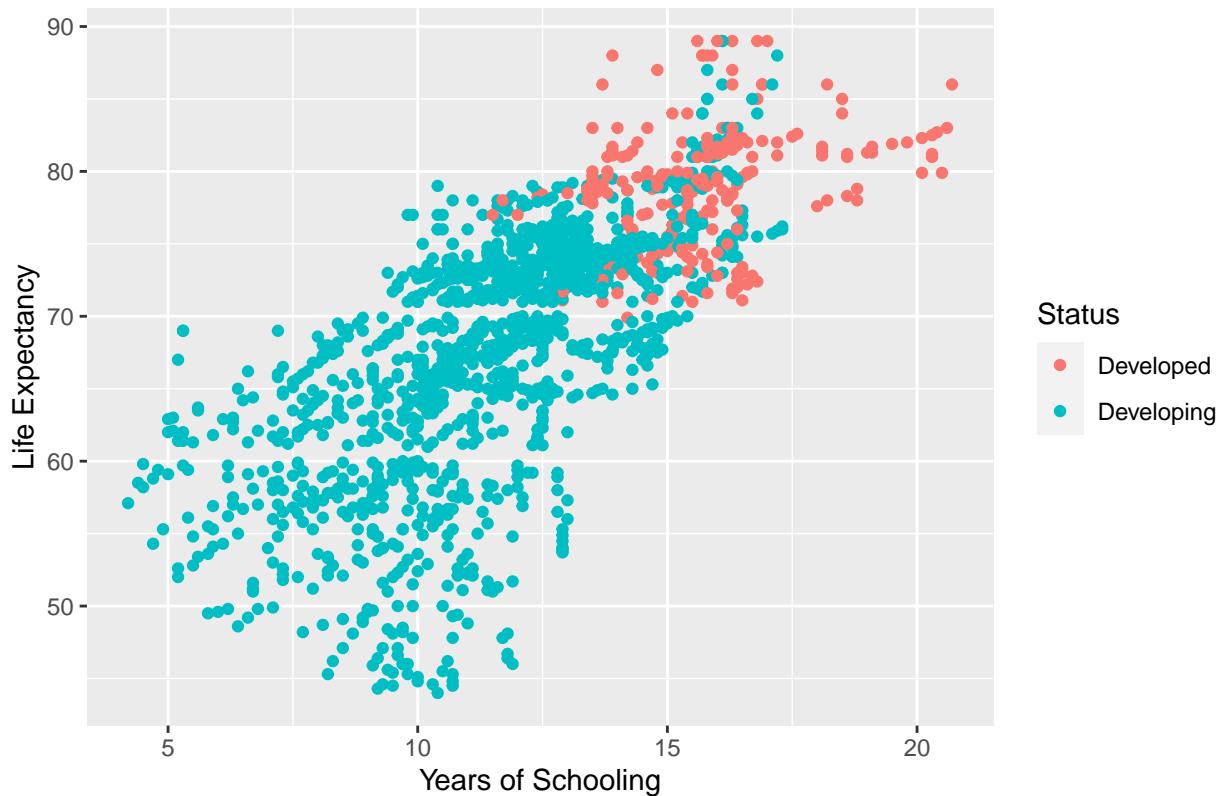
```
# human development index vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Income.composition.of.resources, y=Life.expectancy, color>Status)) +
  ggtitle("Human Development Index vs. Life Expectancy") +
  xlab("Human Development Index") +
  ylab("Life Expectancy")
```

## Human Development Index vs. Life Expectancy



```
# years of schooling vs. life expectancy
ggplot(data=lifeExpectancy) +
  geom_point(mapping=aes(x=Schooling, y=Life.expectancy, color>Status)) +
  ggtitle("Schooling vs. Life Expectancy") +
  xlab("Years of Schooling") +
  ylab("Life Expectancy")
```

## Schooling vs. Life Expectancy



## EDA for non-continuous variables

```
# status vs. life expectancy
developing_expectancy <- lifeExpectancy %>% filter(Status == "Developing") %>% select(Life.expectancy)
developed_expectancy <- lifeExpectancy %>% filter(Status == "Developed") %>% select(Life.expectancy)

boxplot_developing <- ggplot(data=developing_expectancy) +
  geom_boxplot(mapping=aes(x=Life.expectancy)) +
  ggtitle("Life Expectancy for Developing Countries") +
  xlab("Life Expectancy")

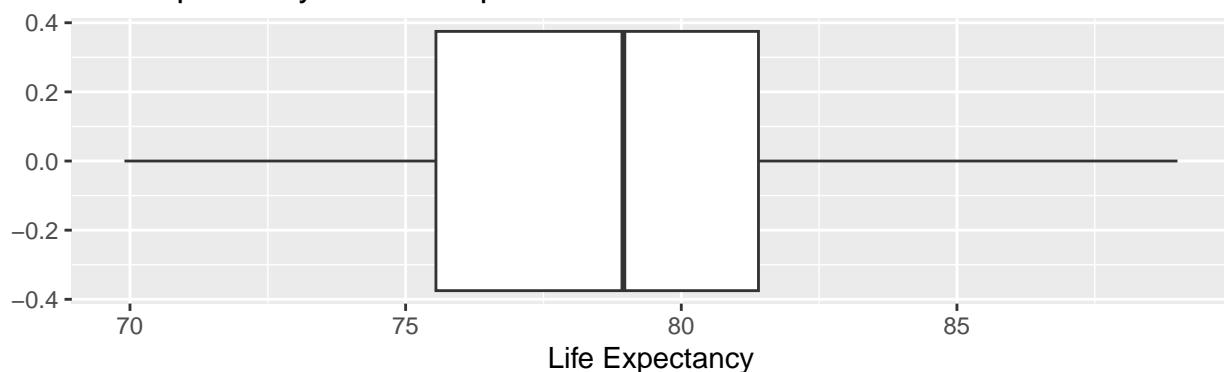
boxplot_developed <- ggplot(data=developed_expectancy) +
  geom_boxplot(mapping=aes(x=Life.expectancy)) +
  ggtitle("Life Expectancy for Developed Countries") +
  xlab("Life Expectancy")

grid.arrange(boxplot_developing, boxplot_developed)
```

## Life Expectancy for Developing Countries



## Life Expectancy for Developed Countries



## Initial modeling using ‘geeglm’ package

### Model 1 (Independent correlation structure, gaussian) Predictors:

- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```
# convert country to factor
lifeExpectancy$Country <- as.factor(lifeExpectancy$Country)

# convert HDI from 0-1 to 0-100 for better interpretability of model estimates
lifeExpectancy$Income.composition.of.resources <- lifeExpectancy$Income.composition.of.resources * 100

# fit initial model
initialModel <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI
+ HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
+ Income.composition.of.resources + Schooling, id = Country,
data = lifeExpectancy, family = "gaussian")
```

```

summary(initialModel)$coefficients

##                               Estimate     Std.err      Wald
## (Intercept)            358.66456721 75.453919934 22.5950515
## StatusDeveloping       -1.36926240  0.819073000  2.7946546
## Year                  -0.15247061  0.038030626 16.0733104
## Adult.Mortality       -0.01748929  0.002305808 57.5305255
## Alcohol                -0.18849973  0.085152151  4.9003850
## BMI                   0.02743665  0.010028907  7.4843649
## HIV.AIDS              -0.44441915  0.049598995 80.2859892
## log(GDP)               0.49802463  0.095407183 27.2483054
## thinness..1.19.years   -0.03846752  0.068831247  0.3123324
## thinness.5.9.years     -0.03477710  0.057004658  0.3721912
## Income.composition.of.resources 0.11102774  0.027585366 16.1996464
## Schooling              0.93039398  0.164142456 32.1286090
##                               Pr(>|W|)
## (Intercept)           1.999991e-06
## StatusDeveloping      9.457914e-02
## Year                  6.093684e-05
## Adult.Mortality       3.330669e-14
## Alcohol                2.685071e-02
## BMI                   6.223702e-03
## HIV.AIDS              0.000000e+00
## log(GDP)              1.789319e-07
## thinness..1.19.years   5.762524e-01
## thinness.5.9.years     5.418123e-01
## Income.composition.of.resources 5.700475e-05
## Schooling              1.442967e-08

# plot residuals
initialResid <- as.vector(initialModel$residuals)
mean_abs_resid_initial <- mean(abs(initialResid))

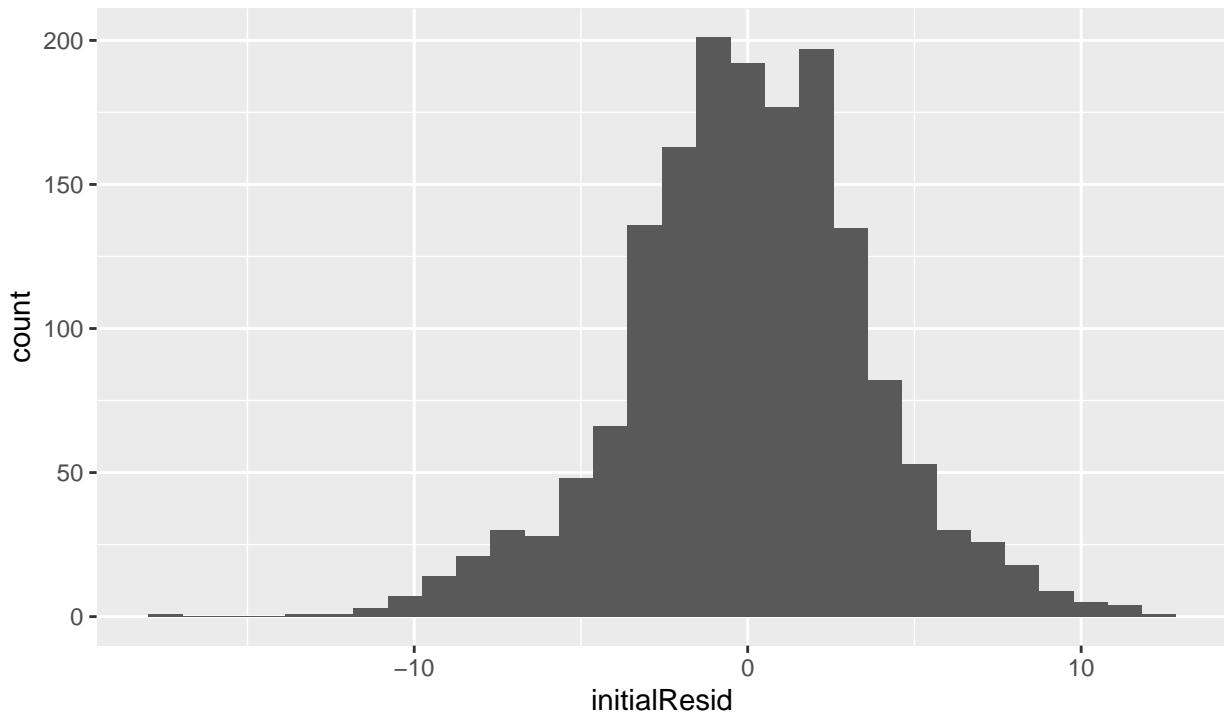
ggplot() +
  geom_histogram(aes(initialResid)) +
  ggtitle(label = "Residuals of initial model \n(independent correlation structure, gaussian)",
           subtitle = paste("Mean absolute error: ", mean_abs_resid_initial))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

Residuals of initial model  
(independent correlation structure, gaussian)

Mean absolute error: 2.81504409706773



**Model 2 (Independent correlation structure) Predictors:**

(Backward selection done using Robust Z)

- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```
# fit second model
secondModel <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI +
HIV.AIDS + log(GDP) + Income.composition.of.resources + Schooling,
id = Country, data = lifeExpectancy, family = "gaussian")
summary(secondModel)$coefficients
```

	Estimate	Std.err	Wald	Pr(> W )
## (Intercept)	362.17234816	74.85685334	23.408169	1.310210e-06
## StatusDeveloping	-1.37665316	0.82975532	2.752637	9.709416e-02
## Year	-0.15465680	0.03767976	16.846968	4.051801e-05
## Adult.Mortality	-0.01753501	0.00229392	58.432646	2.098322e-14
## Alcohol	-0.17648021	0.08527053	4.283454	3.848506e-02
## BMI	0.03410873	0.01044020	10.673663	1.086717e-03
## HIV.AIDS	-0.44699537	0.04902540	83.131154	0.000000e+00
## log(GDP)	0.49770594	0.09523882	27.309744	1.733357e-07

```

## Income.composition.of.resources   0.11214319  0.02707574 17.154775 3.445422e-05
## Schooling                      0.94413800  0.15991610 34.856724 3.548818e-09

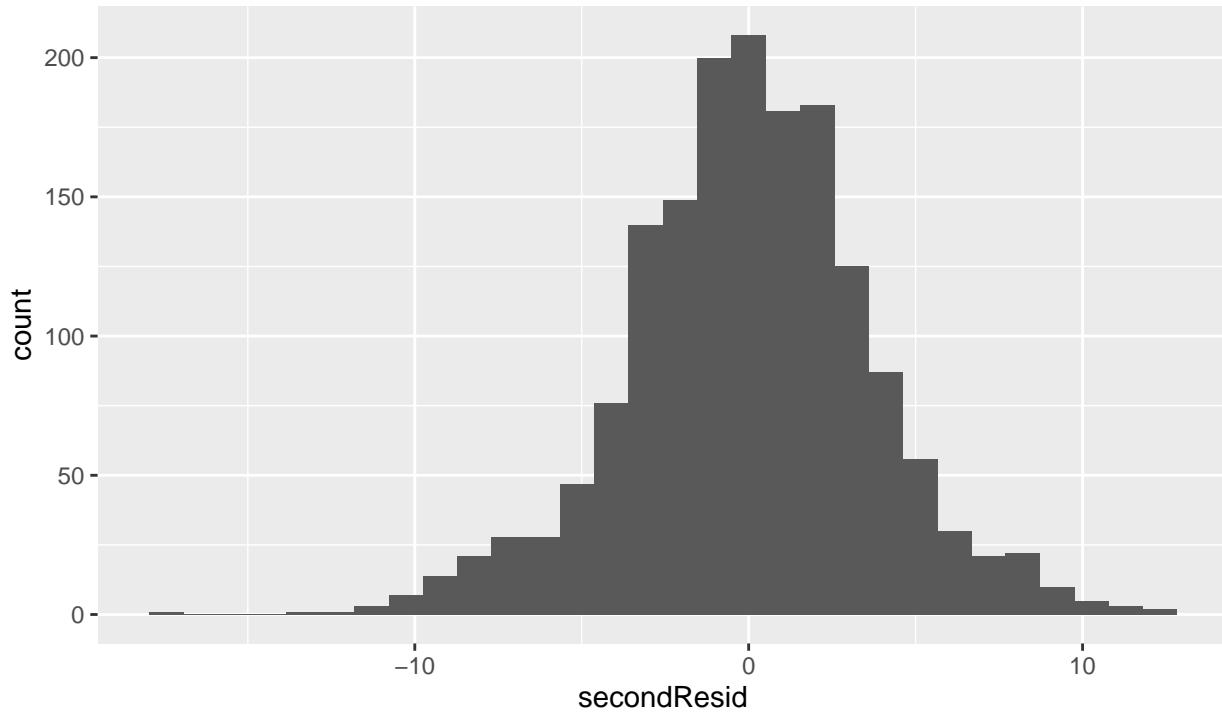
# plot residuals
secondResid <- as.vector(secondModel$residuals)
mean_abs_resid_second <- mean(abs(secondResid))

ggplot() +
  geom_histogram(aes(secondResid)) +
  ggtitle(label = "Residuals of second model \n(independent correlation structure)",
           subtitle = paste("Mean absolute error: ", mean_abs_resid_second))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

**Residuals of second model  
(independent correlation structure)**  
Mean absolute error: 2.82039603325064



### “AR1” correlation

#### Model 1 (AR1 correlation structure) Predictors:

- Status (Developed/Developing)
- Adult mortality rate
- Alcohol intake
- BMI
- HIV/AIDS
- log(GDP)
- Prevalence of thinness among ages 10-19
- Prevalence in thinness among ages 5-9
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```

# fit initial model
initialModel_ar1 <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol + BMI
                           + HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
                           + Income.composition.of.resources + Schooling, id = Country,
                           data = lifeExpectancy, corstr = "ar1", family = "gaussian")
summary(initialModel_ar1)$coefficients

##                                     Estimate      Std.err      Wald
## (Intercept)                 -8.615082e+01 5.706060e+01 2.279534421
## StatusDeveloping            -5.824818e+00 1.040977e+00 31.309973394
## Year                         7.304077e-02 2.906636e-02 6.314649336
## Adult.Mortality             -7.352253e-04 5.368132e-04 1.875834428
## Alcohol                      -5.622125e-03 2.282424e-02 0.060674809
## BMI                          1.463369e-03 2.999144e-03 0.238074626
## HIV.AIDS                     -4.078279e-01 5.870129e-02 48.267915457
## log(GDP)                     1.952508e-03 3.141030e-02 0.003864038
## thinness..1.19.years          -4.008282e-02 3.277668e-02 1.495500857
## thinness.5.9.years            1.521806e-02 2.768839e-02 0.302080819
## Income.composition.of.resources 1.683596e-02 4.707713e-03 12.789572690
## Schooling                    1.103759e+00 1.234923e-01 79.885624828
##                                     Pr(>|W|)
## (Intercept)                   1.310912e-01
## StatusDeveloping              2.199468e-08
## Year                          1.197444e-02
## Adult.Mortality               1.708083e-01
## Alcohol                       8.054325e-01
## BMI                           6.256002e-01
## HIV.AIDS                      3.717804e-12
## log(GDP)                      9.504343e-01
## thinness..1.19.years           2.213649e-01
## thinness.5.9.years              5.825809e-01
## Income.composition.of.resources 3.485567e-04
## Schooling                     0.000000e+00

# plot residuals
initialResid_ar1 <- as.vector(initialModel_ar1$residuals)
mean_abs_resid_initial_ar1 <- mean(abs(initialResid_ar1))

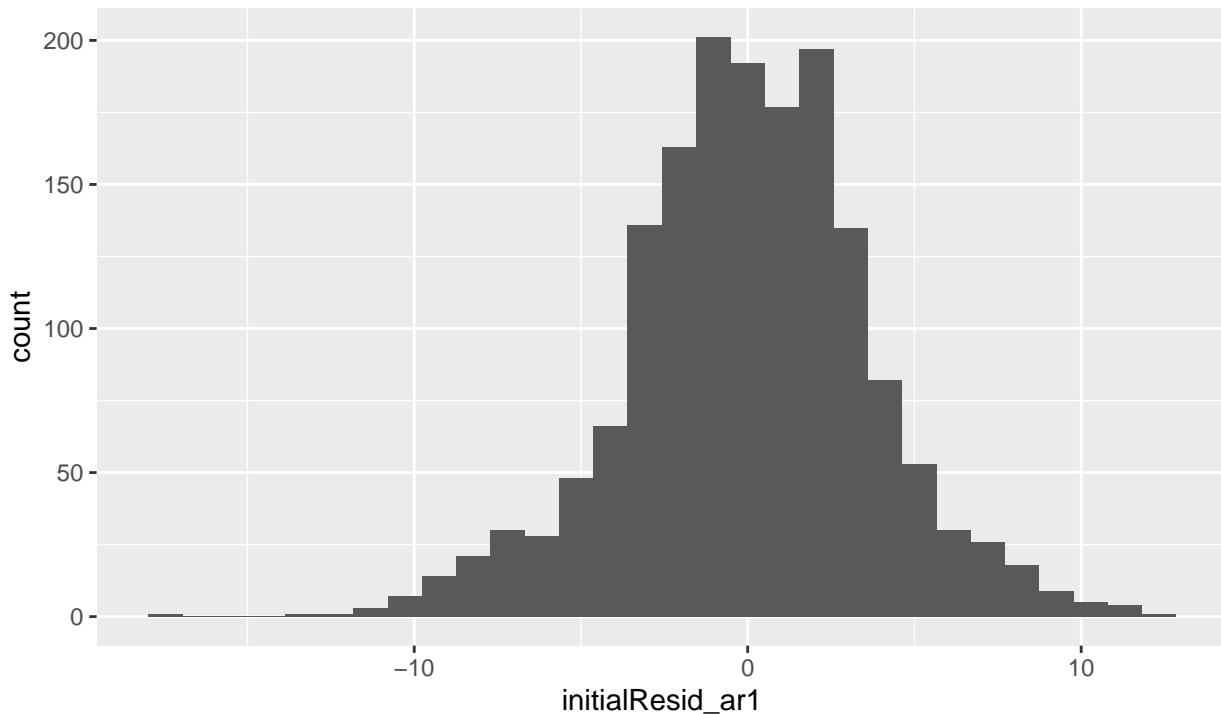
ggplot() +
  geom_histogram(aes(initialResid_ar1)) +
  ggtitle(label = "Residuals of initial model \n(AR1 correlation structure)",
         subtitle = paste("Mean absolute error: ", mean_abs_resid_initial_ar1))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

Residuals of initial model  
(AR1 correlation structure)

Mean absolute error: 2.81504409706773



Using AR1 Correlation structure, perform backwards selection of predictors

**Model 2 (AR1 Correlation Structure) Predictors:**

- Status (Developed/Developing)
- Year
- HIV/AIDS
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```
# fit initial model
model2_ar1 <- geeglm(formula = Life.expectancy ~ Status + Year + HIV.AIDS
+ Income.composition.of.resources + Schooling, id = Country,
  data = lifeExpectancy, corstr = "ar1", family = "gaussian")
summary(model2_ar1)$coefficients
```

	Estimate	Std.err	Wald
## (Intercept)	-92.69412291	56.417504072	2.699459
## StatusDeveloping	-5.95221369	1.031812649	33.277856
## Year	0.07619169	0.028783899	7.006739
## HIV.AIDS	-0.40892008	0.060509651	45.669640
## Income.composition.of.resources	0.01685699	0.004582926	13.529264
## Schooling	1.11410328	0.122475875	82.746532
##		Pr(> W )	
## (Intercept)	1.003823e-01		
## StatusDeveloping	7.988718e-09		
## Year	8.120345e-03		
## HIV.AIDS	1.399758e-11		

```

## Income.composition.of.resources 2.348722e-04
## Schooling 0.000000e+00
# plot residuals
model2Resid_ar1 <- as.vector(model2_ar1$residuals)
mean_abs_resid_model2_ar1 <- mean(abs(model2Resid_ar1))

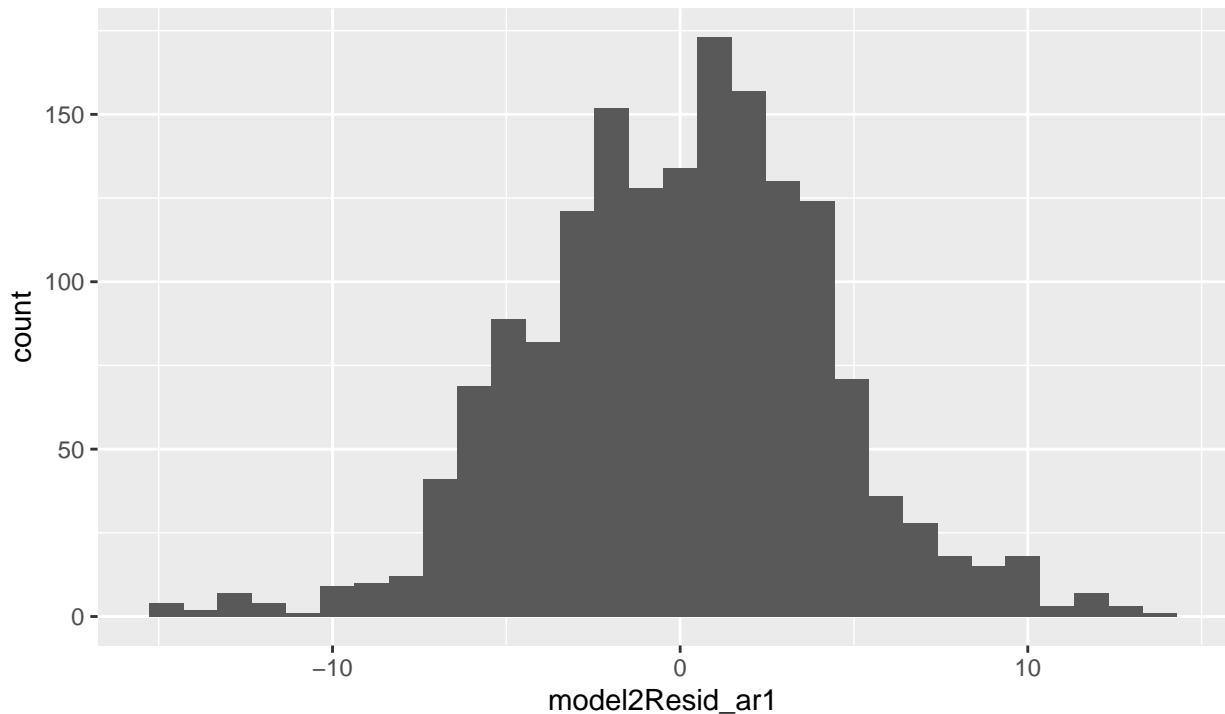
ggplot() +
  geom_histogram(aes(model2Resid_ar1)) +
  ggtitle(label = "Residuals of Model 2 \n(AR1 correlation structure)",
           subtitle = paste("Mean absolute error: ", mean_abs_resid_model2_ar1))

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

### Residuals of Model 2 (AR1 correlation structure)

Mean absolute error: 3.32088500137472



### Compare Independent vs. AR-1 Standard Error:

(Same predictors in both models)

- Status (Developed/Developing)
- Adult mortality rate
- HIV/AIDS
- Human Development Index (in terms of income composition of resources)
- Number of years of schooling

```

# fit independent model w/ same predictors as final AR-1 model to compare standard errors
testIndepModel <- geeglm(formula = Life.expectancy ~ Status + Year + HIV.AIDS
                           + Income.composition.of.resources + Schooling,
                           id = Country, data = lifeExpectancy, family = "gaussian")

```

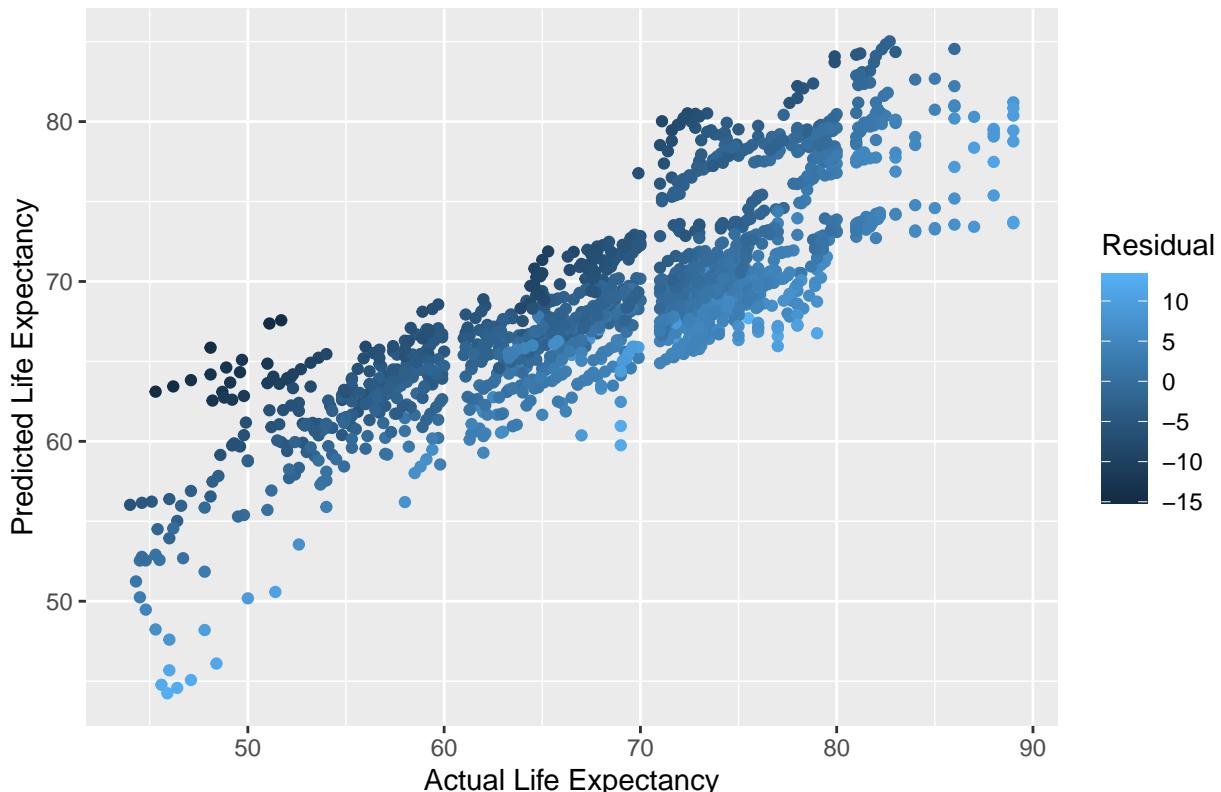
```
summary(testIndepModel)$coefficients
```

	Estimate	Std. err	Wald	Pr(> W )
## (Intercept)	360.5058009	84.26742318	18.302275	1.884819e-05
## StatusDeveloping	-1.2318027	0.97380582	1.600065	2.058940e-01
## Year	-0.1556290	0.04253417	13.387679	2.532826e-04
## HIV.AIDS	-0.6413125	0.08679507	54.594540	1.481038e-13
## Income.composition.of.resources	0.1434149	0.03599286	15.876533	6.761189e-05
## Schooling	1.1997387	0.20442289	34.444057	4.386801e-09

## Using AR-1 Model 2, Examine Predicted vs. Actual Values

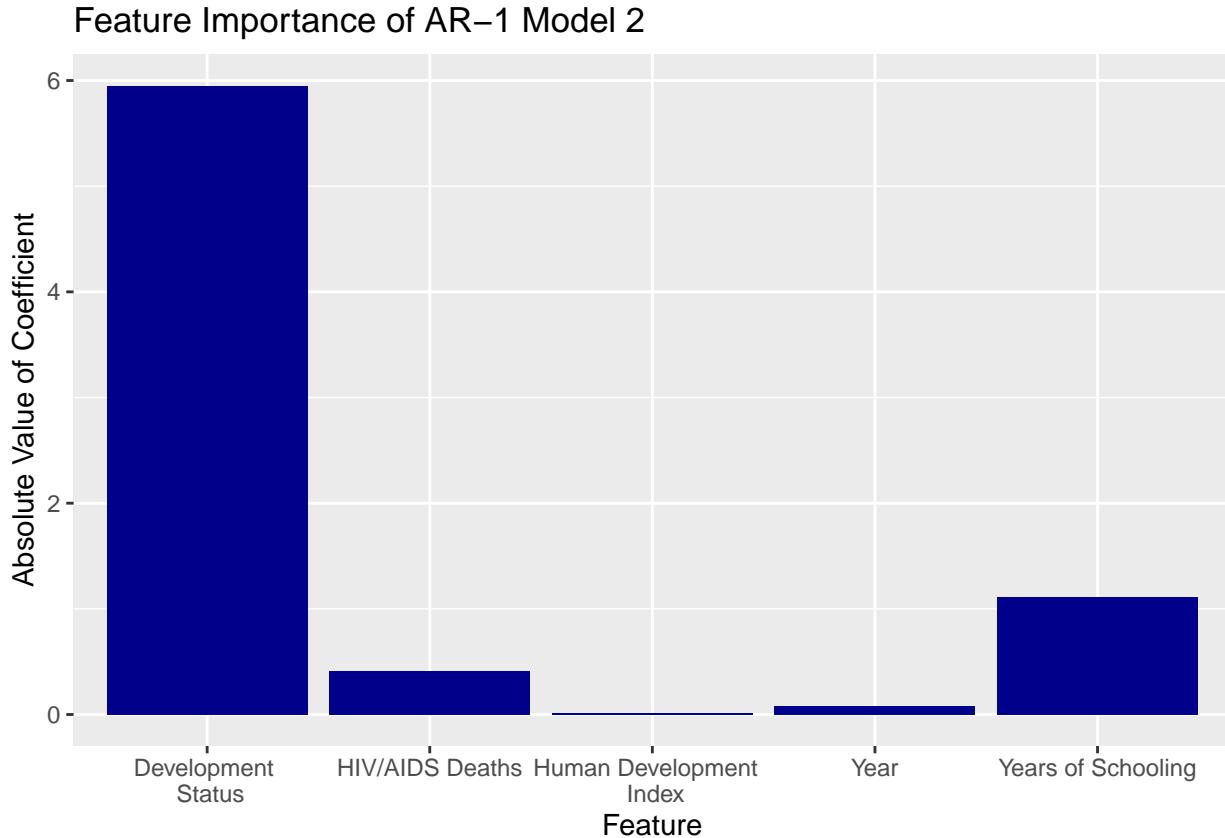
```
fitted_vals <- model2_ar1$fitted.values  
actual_vals <- lifeExpectancy$Life.expectancy  
actual_fitted_df <- as.data.frame(cbind(fitted_vals, actual_vals, model2Resid_ar1))  
names(actual_fitted_df)[1] <- "fitted_vals"  
names(actual_fitted_df)[3] <- "Residual"  
  
ggplot(actual_fitted_df) +  
  geom_point(aes(x=actual_vals, fitted_vals, color=Residual)) +  
  ggtitle("Actual vs. Fitted Values for AR-1 Model 2") +  
  xlab("Actual Life Expectancy") +  
  ylab("Predicted Life Expectancy")
```

Actual vs. Fitted Values for AR-1 Model 2



## Feature Importance in AR-1 Model 2

```
feature_names <- c("Development \n Status", "Year", "HIV/AIDS Deaths", "Human Development \n Index", "Years of Schooling")  
abs_coeff <- abs(model2_ar1$coefficients[-1])  
  
ggplot() +  
  geom_col(aes(x=feature_names, y=abs_coeff), fill = "dark blue") +  
  ggtitle("Feature Importance of AR-1 Model 2") +  
  xlab("Feature") +  
  ylab("Absolute Value of Coefficient")
```



## Exploring new clustering based on similarity of countries

K-Means to create new country clusters

```
kmeans_vars <- continuous_vars[, c(-1, -2, -18)]  
  
k_vec <- 3:30  
tot_within_SS <- c()  
  
# Cross-Validate to find the best K, which minimizes the total within cluster sum of squares  
for (k in k_vec) {  
  
  clustering_k <- kmeans(kmeans_vars, centers=k, nstart=10, iter.max=100)  
  tot_within_SS <- c(tot_within_SS, clustering_k$tot.withinss)
```

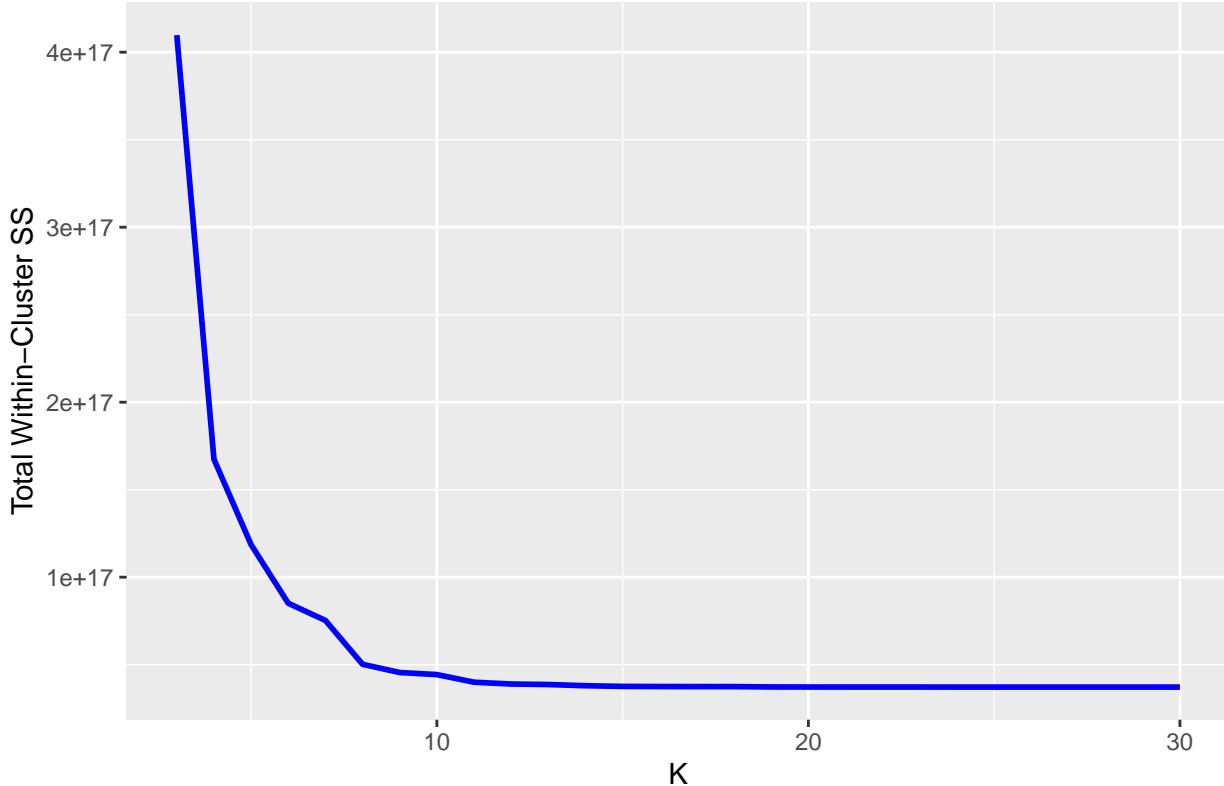
```

}

# plot K vs. total within-cluster SS
ggplot() +
  geom_line(aes(x=k_vec, y=tot_within_SS), color="blue", lwd = 1) +
  ggtitle("Total Within-Cluster Sum of Squares for K = (3, 4, ..., 30)") +
  xlab("K") +
  ylab("Total Within-Cluster SS")

```

Total Within-Cluster Sum of Squares for K = (3, 4, ..., 30)



Optimal number of clusters: K = 10

```

# perform final clustering
kmeans10 <- kmeans(kmeans_vars, centers=10, nstart=10)
new_clusters <- factor(kmeans10$cluster)

```

Compare human development index values for cluster members

```

HDI <- lifeExpectancy$Income.composition.of.resources

# get countries belonging to each cluster
lifeExpectancy$cluster <- new_clusters
countries_cluster <- lifeExpectancy %>% group_by(cluster) %>% summarise(countries = str_c(unique(Country))

ggplot() +
  geom_boxplot(aes(x=new_clusters, y=HDI, fill=new_clusters)) +

```

```
ggtitle("Human Development Index of New Cluster Members") +
  xlab("Cluster")
```



Countries in Cluster 1: Algeria, Argentina, Colombia, Italy, Kenya, Myanmar, Poland, South Africa, Spain, Uganda, Ukraine

Countries in Cluster 2: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Belgium, Bosnia and Herzegovina, Botswana, Brazil, Cameroon, Canada, Central African Republic, Colombia, Costa Rica, Croatia, Eritrea, Ghana, Greece, Indonesia, Iraq, Ireland, Jamaica, Jordan, Kenya, Latvia, Lebanon, Lesotho, Liberia, Lithuania, Madagascar, Malaysia, Mauritania, Mongolia, Morocco, Mozambique, Myanmar, Namibia, Nepal, Nicaragua, Panama, Papua New Guinea, Peru, Poland, Romania, Senegal, South Africa, Spain, Syrian Arab Republic, Tunisia, Turkmenistan, Uganda, Ukraine, Uruguay, Uzbekistan

Countries in Cluster 3: Ethiopia, France, Germany, Philippines, Thailand, Turkey

Countries in Cluster 4: Afghanistan, Algeria, Angola, Australia, Brazil, Canada, Ghana, Indonesia, Iraq, Madagascar, Malaysia, Morocco, Mozambique, Nepal, Peru, Romania, Uganda

Countries in Cluster 5: Brazil, Indonesia, Pakistan

Countries in Cluster 6: Austria, Azerbaijan, Belarus, Belgium, Benin, Bulgaria, Burundi, Chad, Dominican Republic, El Salvador, France, Germany, Greece, Guinea, Honduras, Italy, Jordan,

Mexico, Myanmar, Nicaragua, Papua New Guinea, Paraguay, Philippines, Rwanda, Senegal, Serbia, Sierra Leone, South Africa, Sweden, Tajikistan, Thailand, Togo, Tunisia, Turkey

Countries in Cluster 7: Bangladesh, India, Mexico, Nigeria, Russian Federation

Countries in Cluster 8: Afghanistan, Bangladesh, Brazil, Burkina Faso, Cambodia, Cameroon, Chad, Chile, Ecuador, Ghana, Guatemala, India, Indonesia, Kazakhstan, Madagascar, Malawi, Mali, Mexico, Mozambique, Netherlands, Niger, Nigeria, Pakistan, Romania, Russian Federation, Senegal, Syrian Arab Republic, Zambia, Zimbabwe

Countries in Cluster 9: Afghanistan, Albania, Algeria, Angola, Argentina, Armenia, Australia, Austria, Azerbaijan, Bangladesh, Belarus, Belgium, Belize, Benin, Bhutan, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Burkina Faso, Burundi, Cabo Verde, Cambodia, Cameroon, Canada, Central African Republic, Chad, Chile, China, Colombia, Comoros, Costa Rica, Croatia, Cyprus, Djibouti, Dominican Republic, Ecuador, El Salvador, Equatorial Guinea, Eritrea, Estonia, Ethiopia, Fiji, France, Gabon, Georgia, Germany, Ghana, Greece, Guatemala, Guinea, Guinea-Bissau, Guyana, Haiti, Honduras, Iraq, Ireland, Israel, Italy, Jamaica, Jordan, Kazakhstan, Kenya, Kiribati, Latvia, Lebanon, Lesotho, Liberia, Lithuania, Luxembourg, Madagascar, Malawi, Malaysia, Maldives, Mali, Malta, Mauritania, Mauritius, Mexico, Mongolia, Montenegro, Morocco, Mozambique, Myanmar, Namibia, Netherlands, Nicaragua, Niger, Nigeria, Pakistan, Panama, Papua New Guinea, Paraguay, Peru, Philippines, Poland, Portugal, Romania, Russian Federation, Rwanda, Samoa, Sao Tome and Principe, Senegal, Serbia, Seychelles, Sierra Leone, Solomon Islands, South Africa, Spain, Sri Lanka, Suriname, Swaziland, Sweden, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Togo, Tonga, Trinidad and Tobago, Tunisia, Turkey, Turkmenistan, Uganda, Ukraine, Uruguay, Uzbekistan, Vanuatu, Zambia, Zimbabwe

Countries in Cluster 10: India

---

## Fit Model 1 using new clusters

Model 1 (Unstructured correlation structure, clustering by K-Means clusters):

```
# fit initial model
initialModel_clusters <- geeglm(formula = Life.expectancy ~ Status + Year + Adult.Mortality + Alcohol +
+ HIV.AIDS + log(GDP) + thinness..1.19.years + thinness.5.9.years
+ Income.composition.of.resources + Schooling, id = cluster,
data = lifeExpectancy, corstr = "unstructured", family = "gaussian")
summary(initialModel_clusters)$coefficients
```

	Estimate	Std.err	Wald
## (Intercept)	412.836364926	3.869385e+02	1.1383400
## StatusDeveloping	-1.473929273	7.888183e-01	3.4913980
## Year	-0.181148442	1.923979e-01	0.8864789
## Adult.Mortality	-0.008920334	2.079018e-03	18.4096550
## Alcohol	-0.236824951	7.681688e-02	9.5047690
## BMI	0.029508709	2.773239e-02	1.1322070
## HIV.AIDS	-0.577117408	6.453359e-02	79.9754470
## log(GDP)	0.316827281	3.834731e-01	0.6826143
## thinness..1.19.years	-0.048453554	6.154738e-02	0.6197723
## thinness.5.9.years	-0.052582088	6.310708e-02	0.6942558
## Income.composition.of.resources	0.123730217	6.308320e-02	3.8470175

```

## Schooling          1.169007353 3.651510e-01 10.2491891
##                  Pr(>|W|)
## (Intercept)       0.2860033650
## StatusDeveloping 0.0616884688
## Year              0.3464332873
## Adult.Mortality   0.0000178153
## Alcohol            0.0020493855
## BMI                0.2873050751
## HIV.AIDS           0.0000000000
## log(GDP)           0.4086879505
## thinness..1.19.years 0.4311318977
## thinness.5.9.years 0.4047205777
## Income.composition.of.resources 0.0498345339
## Schooling          0.0013674470

# plot residuals
initialResid_clusters <- as.vector(initialModel_clusters$residuals)
mean_abs_resid_initial_clusters <- mean(abs(initialResid_clusters))

ggplot() +
  geom_histogram(aes(initialResid_clusters)) +
  ggtitle(label = "Residuals of initial model \n(unstructured correlation structure, clusters from K-Means")
  subtitle = paste("Mean absolute error: ", mean_abs_resid_initial_clusters))

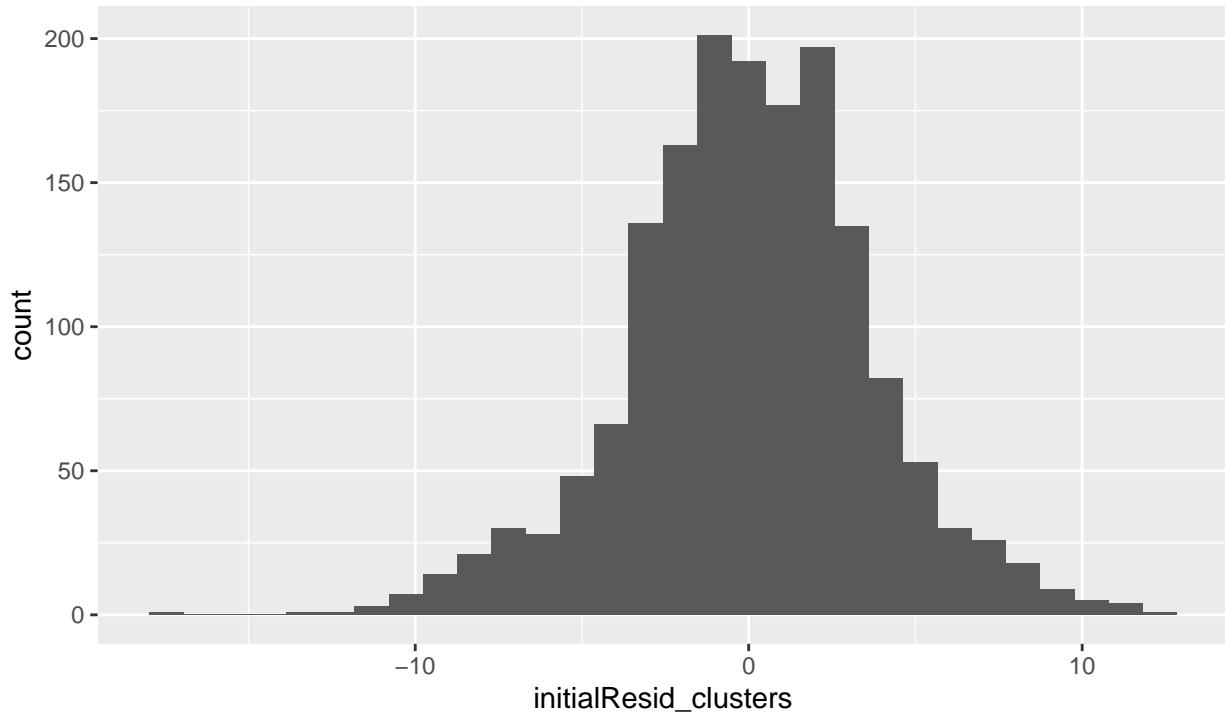
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

```

### Residuals of initial model

(unstructured correlation structure, clusters from K-Means)

Mean absolute error: 2.81504409706773



## Comparing standard errors from clustering by Country vs. K-Means

```
summary(initialModel_ar1)$coefficients
```

	Estimate	Std.err	Wald
## (Intercept)	-8.615082e+01	5.706060e+01	2.279534421
## StatusDeveloping	-5.824818e+00	1.040977e+00	31.309973394
## Year	7.304077e-02	2.906636e-02	6.314649336
## Adult.Mortality	-7.352253e-04	5.368132e-04	1.875834428
## Alcohol	-5.622125e-03	2.282424e-02	0.060674809
## BMI	1.463369e-03	2.999144e-03	0.238074626
## HIV.AIDS	-4.078279e-01	5.870129e-02	48.267915457
## log(GDP)	1.952508e-03	3.141030e-02	0.003864038
## thinness..1.19.years	-4.008282e-02	3.277668e-02	1.495500857
## thinness.5.9.years	1.521806e-02	2.768839e-02	0.302080819
## Income.composition.of.resources	1.683596e-02	4.707713e-03	12.789572690
## Schooling	1.103759e+00	1.234923e-01	79.885624828
##	Pr(> W )		
## (Intercept)	1.310912e-01		
## StatusDeveloping	2.199468e-08		
## Year	1.197444e-02		
## Adult.Mortality	1.708083e-01		
## Alcohol	8.054325e-01		
## BMI	6.256002e-01		
## HIV.AIDS	3.717804e-12		
## log(GDP)	9.504343e-01		
## thinness..1.19.years	2.213649e-01		
## thinness.5.9.years	5.825809e-01		
## Income.composition.of.resources	3.485567e-04		
## Schooling	0.000000e+00		

```
summary(initialModel_clusters)$coefficients
```

	Estimate	Std.err	Wald
## (Intercept)	412.836364926	3.869385e+02	1.1383400
## StatusDeveloping	-1.473929273	7.888183e-01	3.4913980
## Year	-0.181148442	1.923979e-01	0.8864789
## Adult.Mortality	-0.008920334	2.079018e-03	18.4096550
## Alcohol	-0.236824951	7.681688e-02	9.5047690
## BMI	0.029508709	2.773239e-02	1.1322070
## HIV.AIDS	-0.577117408	6.453359e-02	79.9754470
## log(GDP)	0.316827281	3.834731e-01	0.6826143
## thinness..1.19.years	-0.048453554	6.154738e-02	0.6197723
## thinness.5.9.years	-0.052582088	6.310708e-02	0.6942558
## Income.composition.of.resources	0.123730217	6.308320e-02	3.8470175
## Schooling	1.169007353	3.651510e-01	10.2491891
##	Pr(> W )		
## (Intercept)	0.2860033650		
## StatusDeveloping	0.0616884688		
## Year	0.3464332873		
## Adult.Mortality	0.0000178153		
## Alcohol	0.0020493855		
## BMI	0.2873050751		
## HIV.AIDS	0.0000000000		
## log(GDP)	0.4086879505		

```
## thinness..1.19.years          0.4311318977
## thinness.5.9.years           0.4047205777
## Income.composition.of.resources 0.0498345339
## Schooling                      0.0013674470
```