Course 2 Section 3.12 - BUILDING MANY MODELS: FITTING (PLOT ALL THE MODELS)

Jiaying Wu

17/10/2020

```
#load library
library(tidyverse)
library(gapminder)
library(broom)
```

```
# Create variable year1952
gapminder2 <- gapminder %>%
  mutate(year1952 = year-1952)

# Group by country, nest, use mutate to fit a model for each country
by_country <- gapminder2 %>%
  select(country, year1952, lifeExp, continent) %>%
  group_by(country, continent) %>%
  nest() %>%
  mutate(model = purrr::map(data, ~ lm(lifeExp ~ year1952, data = .))) %>%
  ungroup()

# Unnest the model column but do so in a tidy way that returns the intercept and slope coefficient
country_coefs <- by_country %>%
  mutate(model = map(model, broom::tidy)) %>%
  unnest(model) %>%
  #Wrangle the data - intercept and slope coefficient as columns
  select(country, continent, term, estimate) %>%
  spread(term, estimate) %>%
  rename(intercept = `(Intercept)`)

# Extract the R-squared of each country's fitted linear model
country_fit <- by_country %>%
  mutate(model = map(model, broom::glance)) %>%
  unnest(model)
```

Give it a go!

filter the top 12 fitted models of life expectancy based on R^2 and compare them with the bottom 12 fitted models.

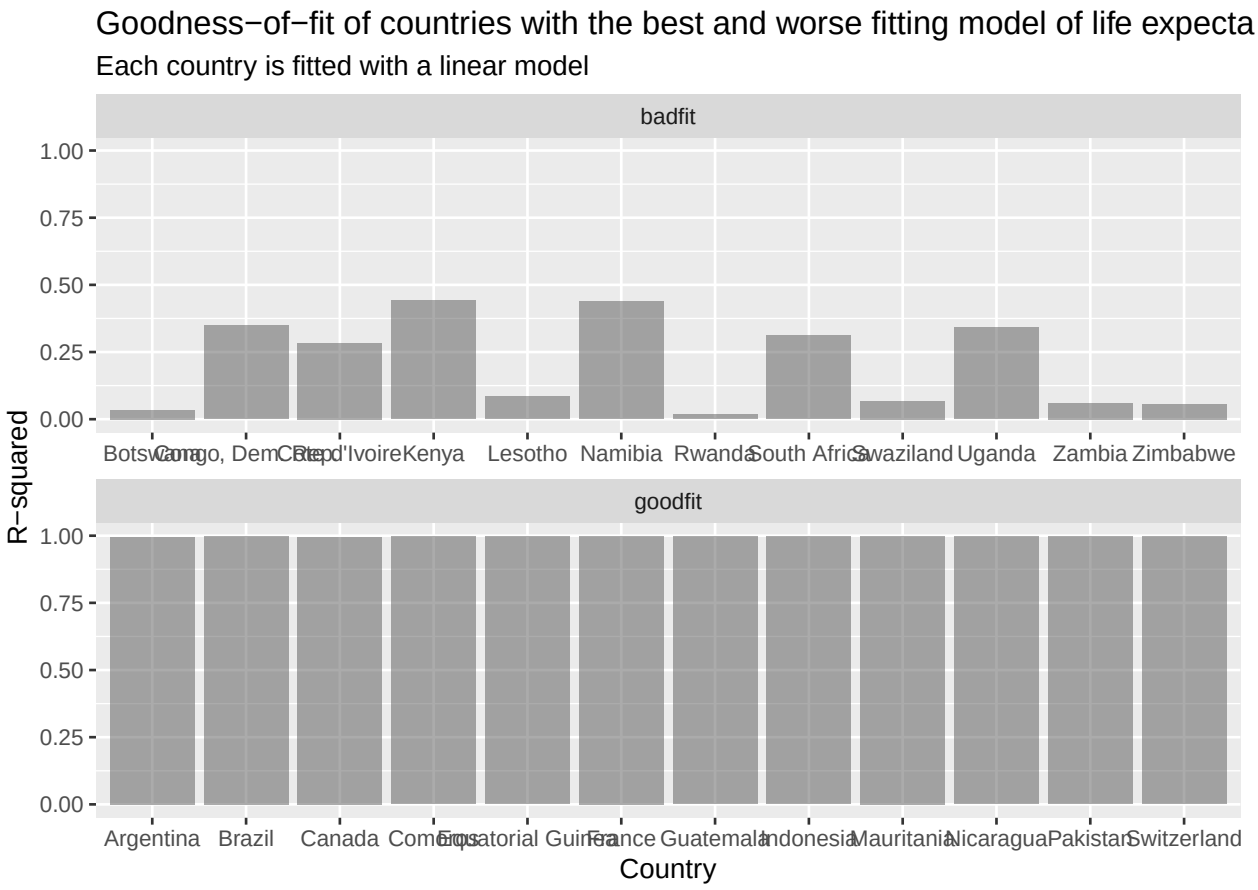
```
# Top and bottom 12 fitted models based on R^2
goodfit <- country_fit %>%
```

```
  top_n(r.squared, n = 12) %>%
  mutate(fit = "goodfit")

badfit <- country_fit %>%
  top_n(r.squared, n = -12) %>%
  mutate(fit = "badfit")

# Row bind goodfit with badfit
good_bad_fit <- bind_rows(goodfit, badfit) %>%
  arrange(desc(r.squared))

# Plot to compare R-squared of country's with good and bad fitting models of life expectancy
good_bad_fit %>%
  ggplot(aes(x = country, y = r.squared)) +
  geom_bar(alpha = 0.5, stat = "identity") +
  facet_wrap(~ fit, nrow = 2, scales = "free_x") +
  labs(title = "Goodness-of-fit of countries with the best and worse fitting model of life expectancy",
       subtitle = "Each country is fitted with a linear model",
       x = "Country",
       y = "R-squared")
```



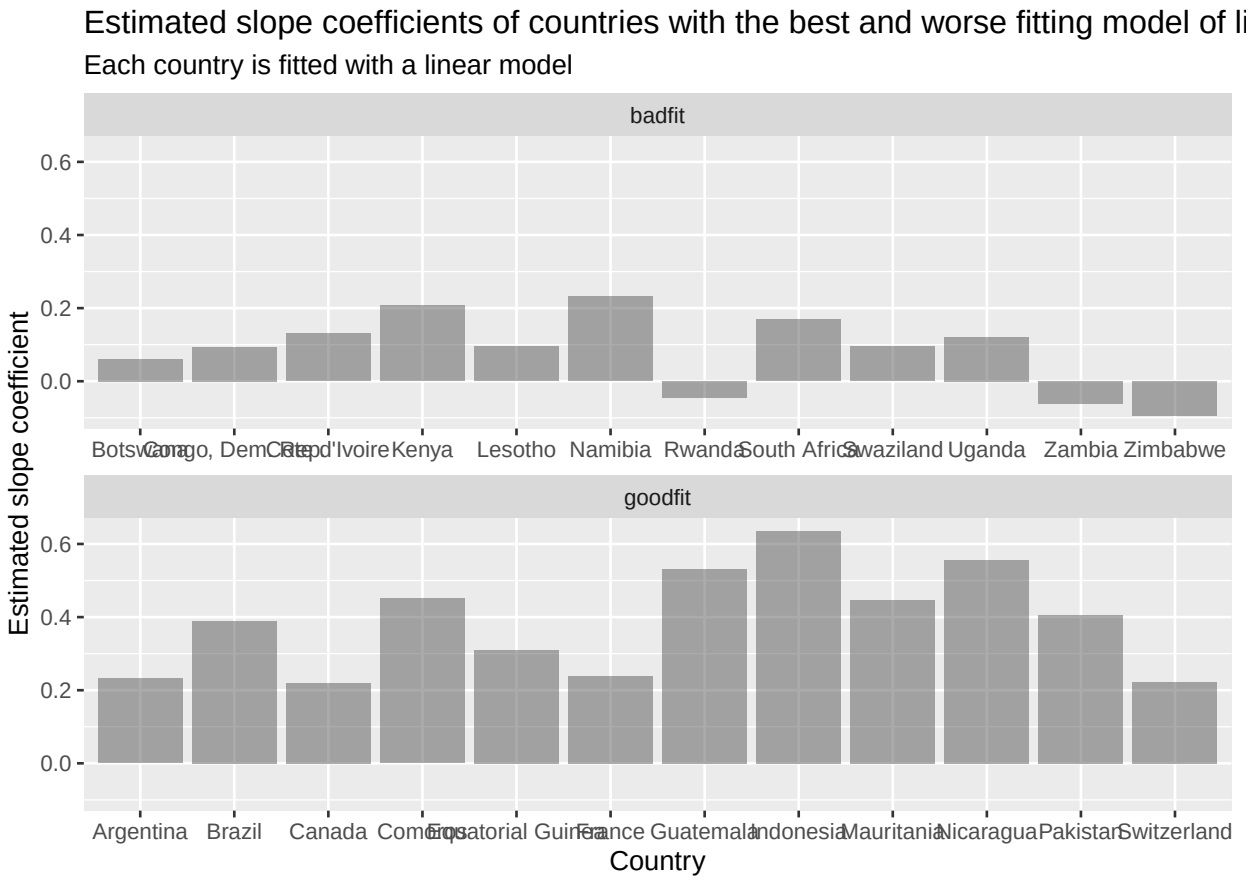
compare the estimated slope coefficient of these country's fitted model of life expectancy.

```r
# Join estimated intercept and slope coefficients to good_bad_fit
good_bad_fit <- left_join(good_bad_fit, (country_coefs %>% select(-continent)), by = "country")

# Print the R-squared and estimated slope coefficients
good_bad_fit %>%
  select(country, fit, r.squared, intercept, year1952)
```

```
## # A tibble: 24 x 5
##    country           fit     r.squared intercept year1952
##    <fct>             <chr>       <dbl>     <dbl>    <dbl>
##  1 Brazil            goodfit     0.998      51.5    0.390
##  2 Mauritania        goodfit     0.998      40.0    0.446
##  3 France            goodfit     0.998      67.8    0.239
##  4 Switzerland       goodfit     0.997      69.5    0.222
##  5 Pakistan          goodfit     0.997      43.7    0.406
##  6 Indonesia         goodfit     0.997      36.9    0.635
##  7 Equatorial Guinea goodfit     0.997      34.4    0.310
##  8 Comoros           goodfit     0.997      40.0    0.450
##  9 Nicaragua         goodfit     0.997      43.0    0.557
## 10 Guatemala         goodfit     0.997      42.1    0.531
## # ... with 14 more rows
```

```r
# Compare the R-squared and estimated slope coefficients for good and bad fitting models
good_bad_fit %>%
  ggplot(aes(x = country, y = year1952)) +
  geom_bar(alpha = 0.5, stat = "identity") +
  facet_wrap(~ fit, nrow = 2, scales = "free_x") +
  labs(title = "Estimated slope coefficients of countries with the best and worse fitting model of life
       subtitle = "Each country is fitted with a linear model",
       x = "Country",
       y = "Estimated slope coefficient")
```



Apply the arrange() function to appropriately arrange columns in country_coefs to answer the following questions. Which country has the:

Q1.lowest improvement in life expectancy?

```
country_coefs %>%
  arrange(year1952)
```

```
## # A tibble: 142 x 4
##    country          continent intercept year1952
##    <fct>            <fct>         <dbl>    <dbl>
##  1 Zimbabwe         Africa         55.2  -0.0930
##  2 Zambia           Africa         47.7  -0.0604
##  3 Rwanda           Africa         42.7  -0.0458
##  4 Botswana         Africa         52.9   0.0607
##  5 Congo, Dem. Rep. Africa         42.0   0.0939
##  6 Swaziland        Africa         46.4   0.0951
##  7 Lesotho          Africa         47.4   0.0956
##  8 Liberia          Africa         39.8   0.0960
##  9 Denmark          Europe         71.0   0.121
## 10 Uganda           Africa         44.3   0.122
## # ... with 132 more rows
```

Q2.highest improvement in life expectancy?

```
country_coefs %>%
  arrange(desc(year1952))
```

```
## # A tibble: 142 x 4
##    country            continent intercept year1952
##    <fct>              <fct>         <dbl>    <dbl>
##  1 Oman               Asia           37.2    0.772
##  2 Vietnam            Asia           39.0    0.672
##  3 Saudi Arabia       Asia           40.8    0.650
##  4 Indonesia          Asia           36.9    0.635
##  5 Libya              Africa         42.1    0.626
##  6 Yemen, Rep.        Asia           30.1    0.605
##  7 West Bank and Gaza Asia           43.8    0.601
##  8 Tunisia            Africa         44.6    0.588
##  9 Gambia             Africa         28.4    0.582
## 10 Jordan             Asia           44.1    0.572
## # ... with 132 more rows
```

Q3.lowest initial life expectancy?

```
country_coefs %>%
  arrange(intercept)
```

```
## # A tibble: 142 x 4
##    country           continent intercept year1952
##    <fct>             <fct>         <dbl>    <dbl>
##  1 Gambia            Africa         28.4    0.582
##  2 Afghanistan       Asia           29.9    0.275
##  3 Yemen, Rep.       Asia           30.1    0.605
##  4 Sierra Leone      Africa         30.9    0.214
##  5 Guinea            Africa         31.6    0.425
##  6 Guinea-Bissau     Africa         31.7    0.272
##  7 Angola            Africa         32.1    0.209
##  8 Mali              Africa         33.1    0.377
##  9 Mozambique        Africa         34.2    0.224
## 10 Equatorial Guinea Africa         34.4    0.310
## # ... with 132 more rows
```

Q4.highest initial life expectancy?

```
country_coefs %>%
  arrange(desc(intercept))
```

```
## # A tibble: 142 x 4
##    country        continent intercept year1952
##    <fct>          <fct>         <dbl>    <dbl>
##  1 Norway         Europe         72.2    0.132
##  2 Iceland        Europe         72.0    0.165
##  3 Netherlands    Europe         71.9    0.137
##  4 Sweden         Europe         71.6    0.166
```

```
##  5 Denmark        Europe         71.0    0.121
##  6 Switzerland    Europe         69.5    0.222
##  7 Canada         Americas       68.9    0.219
##  8 United Kingdom Europe         68.8    0.186
##  9 New Zealand    Oceania        68.7    0.193
## 10 United States  Americas       68.4    0.184
## # ... with 132 more rows
```