Course 2 Section 3.11 - BUILDING MANY MODELS: FITTING

Jiaying Wu

17/10/2020

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

Give it a go!

After fitting the model for each country, choose a country (other than Australia) and share the fitted model with other learners.

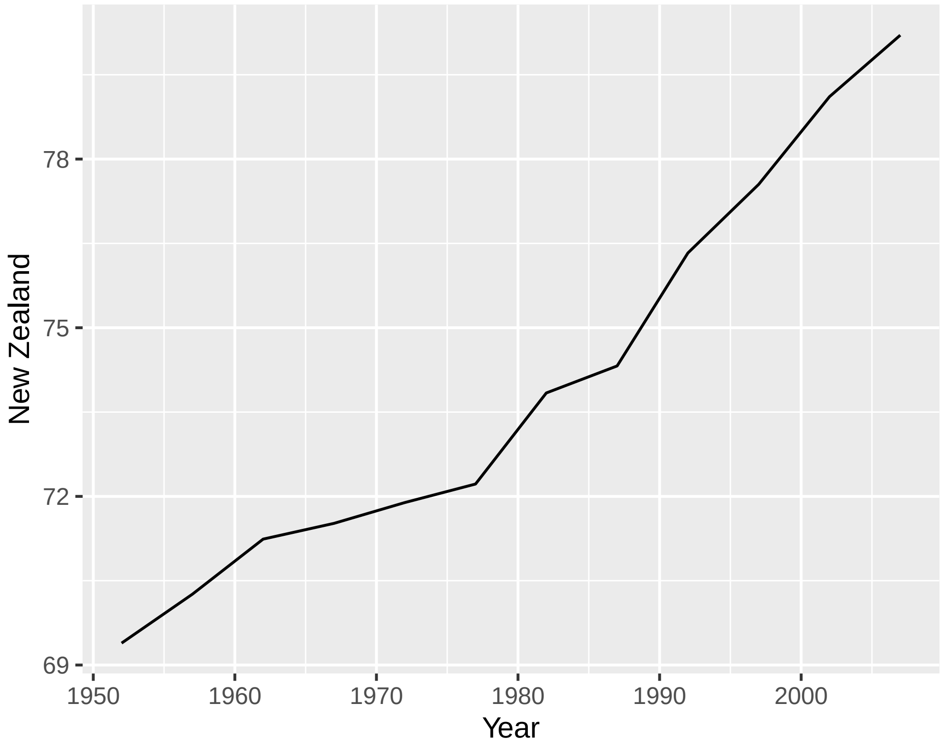
```r
# Mutate year
gapminder2 <- gapminder %>% mutate(year1952 = year-1952)

# Filter for New Zealand only
nz <- gapminder2 %>% filter(country=="New Zealand")

# Look at head of nz
head(nz)
```

```
## # A tibble: 6 x 7
##   country     continent  year lifeExp     pop gdpPercap year1952
##   <fct>       <fct>     <int>   <dbl>   <int>     <dbl>    <dbl>
## 1 New Zealand Oceania    1952    69.4 1994794    10557.        0
## 2 New Zealand Oceania    1957    70.3 2229407    12247.        5
## 3 New Zealand Oceania    1962    71.2 2488550    13176.       10
## 4 New Zealand Oceania    1967    71.5 2728150    14464.       15
## 5 New Zealand Oceania    1972    71.9 2929100    16046.       20
## 6 New Zealand Oceania    1977    72.2 3164900    16234.       25
```

```r
# Line plot of life expectancy in New Zealand over time
ggplot(data=nz, aes(x=year, y=lifeExp)) +
  geom_line() +
  labs(x = "Year", y = "New Zealand")
```



The line plot shows that the increase in life expectancy in New Zealand slowed from 1962 to 1977, which might also related to mortality during the Vietnam war.

```r
# Fit model of life expectancy using year1952 as explanatory variable
nz_lm <- lm(lifeExp~year1952, data = nz)

# Tidy output of fitted model
tidy(nz_lm)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   68.7      0.437      157.  2.66e-18
## 2 year1952       0.193    0.0135      14.3 5.41e- 8
```

$\widehat{lifeExp} = 68.6869 + 0.1928 year1052$

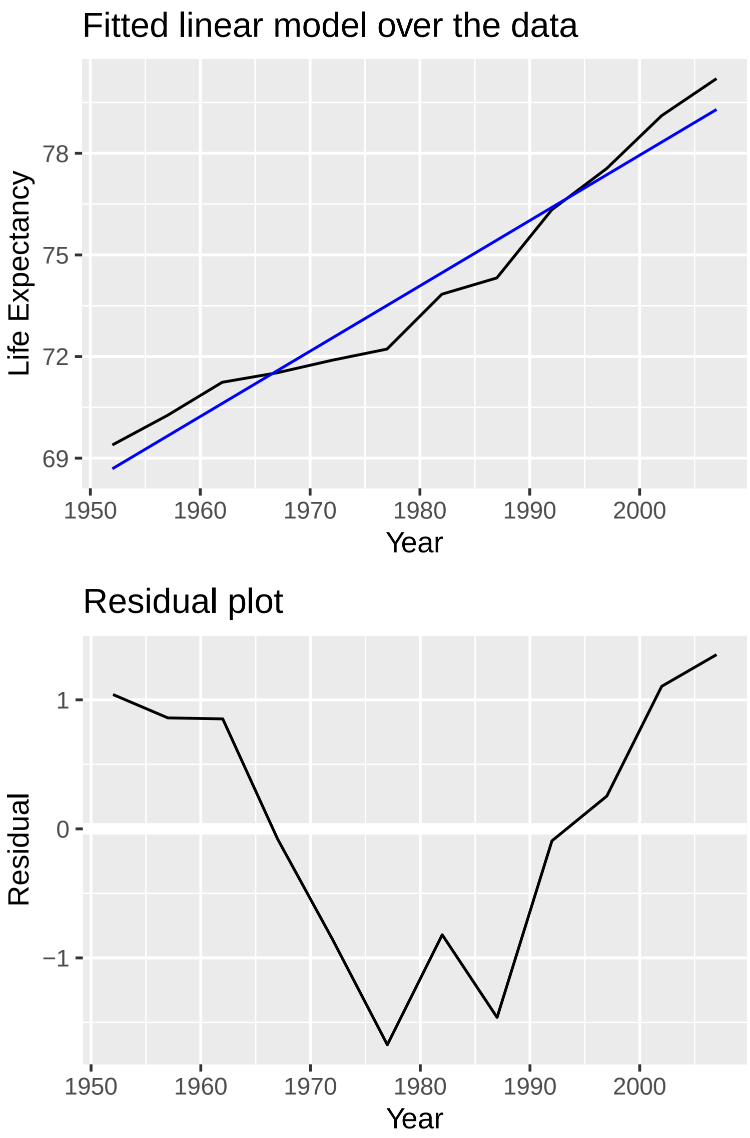
The fitted model estimates that life expectancy in Australia increases by approximately 1.9 years every decade, on average.

```r
# Append fitted and residual value into oz (training data)
nz_mod <- augment(nz_lm, nz)

# Plot fitted model over the data
p1 <- ggplot(data=nz_mod, aes(x=year, y=lifeExp)) +
  geom_line() +
  geom_line(aes(y=.fitted), colour="blue") +
  labs(x = "Year", y = "Life Expectancy", title = "Fitted linear model over the data")
```

```
# Plot of residuals
p2 <- ggplot(data=nz_mod, aes(x=year, y=.std.resid)) +
  geom_hline(yintercept=0, colour="white", size=2) +
  geom_line() +
  labs(x = "Year", y = "Residual", title = "Residual plot")

# Group above plots into grid
gridExtra::grid.arrange(p1, p2)
```


Fitted linear model over the data

Residual plot

The fitted linear model, which estimates that life expectancy in New Zealand increases by a constant amount (approximately 1.9 years each decade), underpredicts from 1967 to 1973, as shown on the residual plot.

Fit all countries

1.Group the data by country and turn it into a nested data frame. This nested data frame will contain a new column, where each row in the column contains country-specific data.

```
# Group by country then nest
by_country <- gapminder2 %>%
  select(country, year1952, lifeExp, continent) %>%
  group_by(country, continent) %>% # Country cannot belong in multiple continents so including this wil
  nest()
```

2.Use map() to apply lm() on each country.

```
# Using mutate to fit a model for each country stores them in by_country
by_country <- by_country %>%
  mutate(
    model = purrr::map(data, ~ lm(lifeExp ~ year1952, data = .))
  )
```

3.Each row in the nested data frame is country-specific i.e. a single row contains the data and fitted linear model of a single country. Since you're interested in each country's fitted model, you will need to unnest the model column. Since each row in the model column contains an lm object, to unnest this requires that the lm objects are first tidied as a tibble. This is done with the tidy() function (recall that when tidy() is used in an lm object, it returns the estimated intercept and slope coefficient(s) in a tibble).

```
# 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)
```

4.Reorganise the data for analysis. Using the spread() function, you can organise the data so that the estimated intercept and slope coefficient corresponding to each country's fitted model are placed in columns.

```
# Wrangle the data - intercept and slope coefficient as columns
country_coefs <- country_coefs %>%
  select(country, continent, term, estimate) %>%
  spread(term, estimate) %>%
  rename(intercept = ‘(Intercept)‘)
```

You have now fitted a linear model of life expectancy for all 142 countries! The code chunk below returns the estimated coefficients of the fitted linear model for 6 countries at the head of the data but you can use the filter verb to choose a country.

```
# Look at the top of country_coefs
head(country_coefs)
```

```
## # A tibble: 6 x 4
## # Groups:   country, continent [6]
##   country     continent intercept year1952
##   <fct>       <fct>         <dbl>    <dbl>
## 1 Afghanistan Asia           29.9    0.275
## 2 Albania     Europe         59.2    0.335
## 3 Algeria     Africa         43.4    0.569
## 4 Angola      Africa         32.1    0.209
## 5 Argentina   Americas       62.7    0.232
## 6 Australia   Oceania        68.4    0.228
```

```
# Filter for the estimated coefficients in the fitted model of AUS
country_coefs %>%
  filter(country == "New Zealand")
```

```
## # A tibble: 1 x 4
## # Groups:   country, continent [1]
##   country     continent intercept year1952
##   <fct>       <fct>         <dbl>    <dbl>
## 1 New Zealand Oceania        68.7    0.193
```

It is also possible to use a for loop to compute the slope and intercept for each country.

```
n <- length(table(gapminder2$country))

country_coefs <- tibble(country=gapminder2$country[seq(1, 1704, 12)],
                   continent=gapminder2$continent[seq(1, 1704, 12)],
                   intercept=rep(0,n),
                   year1952=rep(0,n))

for (i in 1:n) {
  sub <- gapminder2 %>% filter(country==country_coefs$country[i])
  sub_lm <- lm(lifeExp~year1952, data=sub)
  sub_lm_coefs <- coefficients(sub_lm)
  country_coefs$intercept[i] <- sub_lm_coefs[1]
  country_coefs$year1952[i] <- sub_lm_coefs[2]
}

head(country_coefs)
```

```
## # A tibble: 6 x 4
##   country     continent intercept year1952
##   <fct>       <fct>         <dbl>    <dbl>
## 1 Afghanistan Asia           29.9    0.275
## 2 Albania     Europe         59.2    0.335
## 3 Algeria     Africa         43.4    0.569
## 4 Angola      Africa         32.1    0.209
## 5 Argentina   Americas       62.7    0.232
## 6 Australia   Oceania        68.4    0.228
```