Review In-Class

Wells

2023-03-03

Multilinear Regression

```
library(gapminder)
data("gapminder")
```

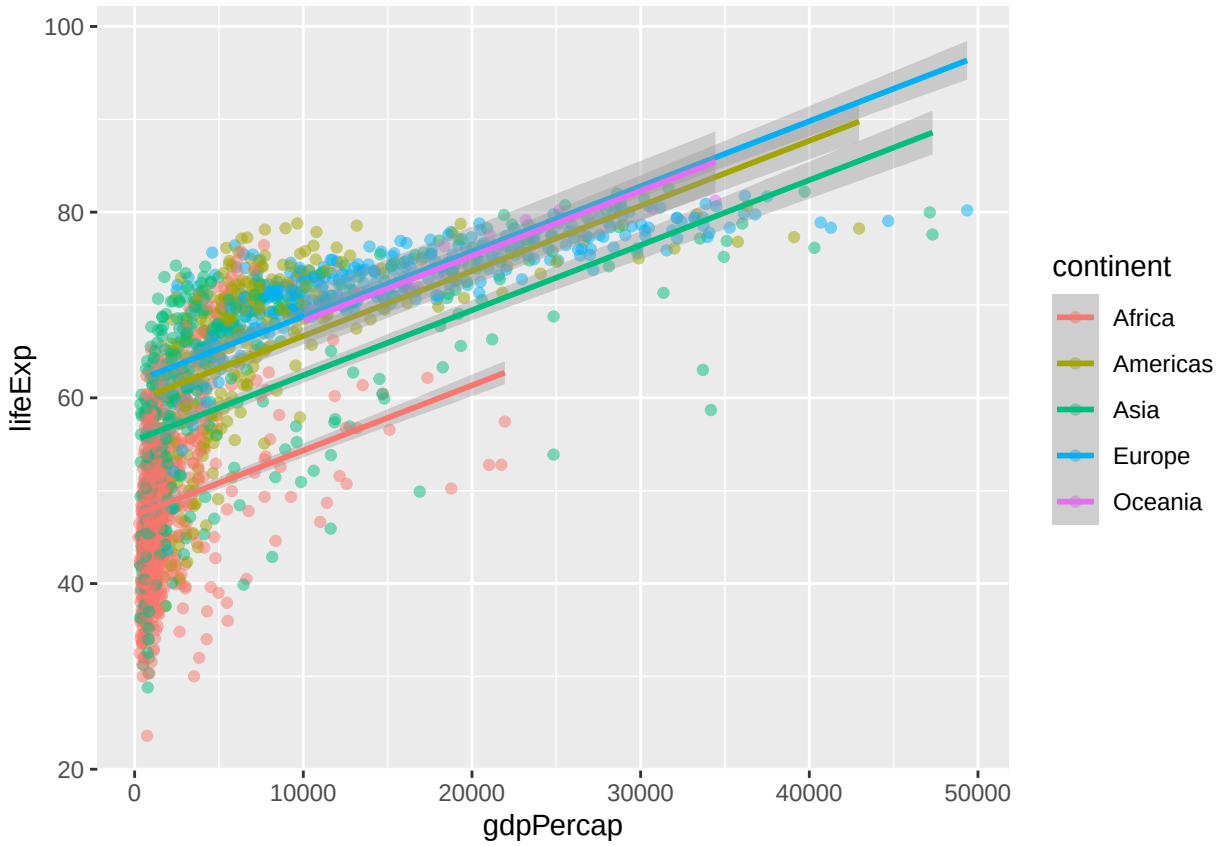
What factors predict life expectancy in a country?

response = lifeExp

possible explanatory

- continent
- year
- population
- gdpPercap

```
filter(gapminder, gdpPercap < 50000) %>%
ggplot(aes( x = gdpPercap, y = lifeExp, color = continent)) +
  geom_point(alpha = .5)+
  geom_parallel_slopes()
```



Make a linear model (or find multiple regression equation) show life expectancy as function of gdpPercap and continent and population.

(this uses + inside the lm)

```
life_mod <- lm(lifeExp ~ gdpPercap + continent + pop, data = gapminder)
get_regression_table(life_mod)
```

```
## # A tibble: 7 x 7
##   term                 estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               47.8      0.34     141.     0        47.1     48.5
## 2 gdpPercap                0        0         19.2    0         0        0
## 3 continent: Americas     13.5      0.6       22.5    0        12.3     14.7
## 4 continent: Asia          8.19     0.571     14.3    0         7.07     9.31
## 5 continent: Europe       17.5      0.625     28.0    0        16.2     18.7
## 6 continent: Oceania      18.1      1.78      10.1    0        14.6     21.6
## 7 pop                      0        0          3.33   0.001     0        0
```

Why is slope 0?

because gdpPercap is on large scale

```
gapminder2 <- gapminder %>%
  mutate(gdp10k = gdpPercap/10000)
```

```
life_mod <- lm(lifeExp ~ gdp10k + continent, data = gapminder2)
get_regression_table(life_mod)
```

```
## # A tibble: 6 x 7
```

```
##   term                 estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept               47.9      0.34      141.        0    47.2     48.6
## 2 gdp10k                   4.45     0.235      18.9       0     3.99     4.91
## 3 continent: Americas     13.6      0.601      22.6       0    12.4     14.8
## 4 continent: Asia          8.66     0.555      15.6       0     7.57     9.75
## 5 continent: Europe       17.6      0.626      28.1       0    16.3     18.8
## 6 continent: Oceania      18.1      1.79       10.2       0    14.6     21.7
```

Life = 47.9 + 4.5 gdpt10k + 13.6 Americas + 8.7 Asia + 17.6 Europe + 18.1 Oceania

What is life expectancy for Americas country with gdp10k of 50k?

Life = 47.9 + 4.5*5 + 13.6

```
47.9 + 4.5*5 + 13.6
```

```
## [1] 84
```

What continent is missing?

```
gapminder %>%
  group_by(continent)%>%
  summarize(number = n())
```

```
## # A tibble: 5 x 2
##   continent number
##   <fct>      <int>
## 1 Africa       624
## 2 Americas     300
## 3 Asia         396
## 4 Europe       360
## 5 Oceania       24
```

```
gapminder %>%
  count(continent)
```

```
## # A tibble: 5 x 2
##   continent     n
##   <fct>     <int>
## 1 Africa      624
## 2 Americas    300
## 3 Asia        396
## 4 Europe      360
## 5 Oceania      24
```

Suppose we have an african country with gdp 50k. What is predicted life exp?

Life = 47.9 + 4.5 gdpt10k + 13.6 Americas + 8.7 Asia + 17.6 Europe + 18.1 Oceania

47.9 + 4.5*5

```
47.9 + 4.5*5
```

```
## [1] 70.4
```

Compare to simple linear model

Find linear model for life expectancy as function of just gdp

```
life_mod_simple <- lm(lifeExp ~ gdp10k, data = gapminder2)
get_regression_table(life_mod_simple)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    54.0      0.315     171.        0    53.3     54.6
## 2 gdp10k        7.65     0.258      29.7       0     7.14     8.15
```

WHy is coefficient on gdp10k different in simple linear regression than multiple linear regression?

7.649 is increase in life expectancy per 1 unit increase in gdp10k.