Homework 5 Bivariate Relationships

Walid Medani

2021-04-10

```
.chunkcolor {
background-color: MistyRose;
}
options(scipen = 9999)
```

Introduction

Due to my previous dataset being limited in showcasing bi-variate relationships, I've chosen the gapminder dataset that explores population, life expectancy, and GDP of all countries from 1952 to 2007.

Below are the summary statistics of the dataset:

```
summary(gapminder) %>%
  kbl() %>%
  kable_material_dark(c("striped", "hover"))
```

	country	continent	year	lifeExp	pop	gdpPercap
	Afghanistan: 12	Africa :624	Min. :1952	Min. :23.60	Min. :6.001e+04	Min. : 241.2
	Albania : 12	Americas:300	1st Qu.:1966	1st Qu.:48.20	1st Qu.:2.794e+06	1st Qu.: 1202.1
	Algeria : 12	Asia :396	Median :1980	Median :60.71	Median :7.024e+06	Median : 3531.8
	Angola : 12	Europe :360	Mean :1980	Mean :59.47	Mean :2.960e+07	Mean : 7215.3
	Argentina : 12	Oceania : 24	3rd Qu.:1993	3rd Qu.:70.85	3rd Qu.:1.959e+07	3rd Qu.: 9325.5
	Australia : 12	NA	Max. :2007	Max. :82.60	Max. :1.319e+09	Max. :113523.1
	(Other) :1632	NA	NA	NA	NA	NA

Countries Cuba and South Africa are randomly selected to observe life expectancy.

```
gapminder %>%
  select(country, lifeExp) %>%
  filter(country == "Cuba" |
         country == "South Africa") %>%
  group_by(country) %>%
  summarise(avglife = mean(lifeExp)) %>%
  kbl() %>%
  kable_material_dark(c("striped", "hover"))
```

Is the difference between Cuba and South Africa's life expectancy due to chance? Since I'm comparing the means of two groups, a t-test

country	avglife
Cuba	71.04508
South Africa	53.99317

can validate the difference in life expectancy and conclude that it isn't due to chance in the sample.

```
southcuba <- gapminder %>%
  select(country, lifeExp) %>%
  filter(country == "Cuba" |
         country == "South Africa")

t.test(data = southcuba, lifeExp ~ country)
```

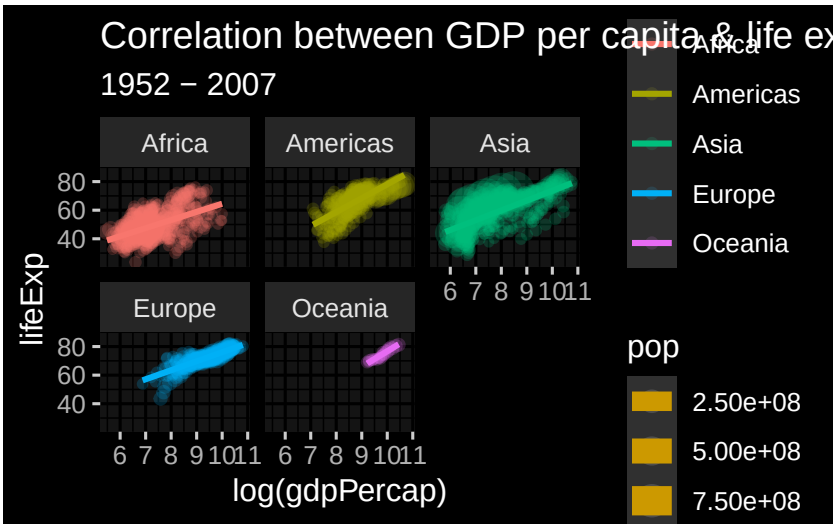
```
##
##  Welch Two Sample t-test
##
## data:  lifeExp by country
## t = 7.2689, df = 21.788, p-value = 2.957e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.18416 21.91967
## sample estimates:
##         mean in group Cuba mean in group South Africa
##                   71.04508                   53.99317
```

Linear Regression

Is there a correlation between life expectancy and gdp per capital and population as seen in the graph below? To find out, I use a multivariate linear regression model by placing GDP & population as the explanatory variables. With p-values below 0.05, our model concludes that there is in fact a correlation between life expectancy and gdp + population.

```
gapminder %>%
  filter(gdpPercap < 50000) %>%
  ggplot(aes(
          x =log(gdpPercap),
          y=lifeExp,
          color=continent,
          size=pop)) +
  geom_point(alpha=0.2)+
  geom_smooth(method = lm)+
```

```
  labs(title = "Correlation between GDP per capita & life expectancy",
       subtitle = "1952 - 2007") +
  theme_fivethirtyeight()+
  ggdark::dark_mode()+
  facet_wrap(~continent)
```




```
summary(lm(lifeExp ~ gdpPercap+pop, gapminder))
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap + pop, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.754  -7.745   2.055   8.212  18.534
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.365e+01  3.225e-01  166.36  < 2e-16 ***
## gdpPercap   7.676e-04  2.568e-05   29.89  < 2e-16 ***
## pop         9.728e-09  2.385e-09    4.08 4.72e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.44 on 1701 degrees of freedom
## Multiple R-squared:  0.3471, Adjusted R-squared:  0.3463
## F-statistic: 452.2 on 2 and 1701 DF,  p-value: < 2.2e-16
```