Final Project

Eric Han, Minghao Sun, Qifan Yang, Steven Luo, Zihan Jin

2024-04-06

Data cleaning

```
country_indicators <-
  read_csv("country_indicators.csv") %>%
  select(-...1) %>%  # remove first column
  select(iso3, everything()) %>%  # reorder the columns to put iso3 as column 1
  rename(country_code_iso3 = iso3)  # rename first column to country_code_iso3
```

```
country_code <- read.csv("country_codes.csv") %>% # read the file
  select(ISO.alpha3.Code..M49., everything()) %>%  # reorder the columns to put iso3 as column 1
  rename(country_code_iso3 = ISO.alpha3.Code..M49.) %>% # rename first column to country_code_iso3
  rename(continent = Region.Name_en..M49.) %>% # rename the column representing continent name
  rename(country_name = Country.or.Area_en..M49.) %>% # rename the column representing country names
  rename(sub_region_name = Sub.region.Name_en..M49.) %>%  # rename the column representing sub-region n
  select(country_code_iso3, country_name, continent, sub_region_name)
```

```
project_data <- inner_join(x=country_code, y=country_indicators, by="country_code_iso3") %>%
  rename(under5_mortality_rate = "sowc_child-mortality__under-five-mortality-rate_2021") %>%
  rename(delivery_care = "sowc_maternal-and-newborn-health__delivery-care-2016-2021-r_skilled-birth-att
  rename(mcv = "sowc_child-health__intervention-coverage_immunization-for-vaccine-preventable-diseases-
  rename(hepb = "sowc_child-health__intervention-coverage_immunization-for-vaccine-preventable-diseases
  rename(hib = "sowc_child-health__intervention-coverage_immunization-for-vaccine-preventable-diseases-
  rename(adolescent_mortality = "sowc_adolescent-health__adolescent-mortality-rate-2021_aged-10-19_tota


  select(country_code_iso3, country_name, continent, sub_region_name, hdr_le_2021, under5_mortality_rat

head(project_data)
```

```
##   country_code_iso3 country_name continent sub_region_name hdr_le_2021
## 1               DZA      Algeria    Africa Northern Africa     76.3767
## 2               EGY        Egypt    Africa Northern Africa     70.2207
## 3               LBY        Libya    Africa Northern Africa     71.9112
## 4               MAR      Morocco    Africa Northern Africa     74.0419
## 5               SDN        Sudan    Africa Northern Africa     65.2667
## 6               TUN      Tunisia    Africa Northern Africa     73.7719
##   under5_mortality_rate delivery_care mcv hepb hib adolescent_mortality
## 1              22.33529          98.8  80   91  91             3.839431
## 2              18.95783          91.5  96   96  96             5.672495
## 3              10.76593          99.9  73   73  73             4.226789
## 4              17.99557          86.6  99   99  99             3.019391
## 5              54.89554          77.7  81   84  84            13.074928
```

```
## 6              16.32134          99.5  95   95  97             4.672385
```

```
data <- project_data %>%
  filter(!is.na(hdr_le_2021))

head(data)
```

```
##   country_code_iso3 country_name continent sub_region_name hdr_le_2021
## 1               DZA      Algeria    Africa Northern Africa     76.3767
## 2               EGY        Egypt    Africa Northern Africa     70.2207
## 3               LBY        Libya    Africa Northern Africa     71.9112
## 4               MAR      Morocco    Africa Northern Africa     74.0419
## 5               SDN        Sudan    Africa Northern Africa     65.2667
## 6               TUN      Tunisia    Africa Northern Africa     73.7719
##   under5_mortality_rate delivery_care mcv hepb hib adolescent_mortality
## 1              22.33529          98.8  80   91  91             3.839431
## 2              18.95783          91.5  96   96  96             5.672495
## 3              10.76593          99.9  73   73  73             4.226789
## 4              17.99557          86.6  99   99  99             3.019391
## 5              54.89554          77.7  81   84  84            13.074928
## 6              16.32134          99.5  95   95  97             4.672385
```

```
q2_indicators <- project_data %>%
  filter(!is.na(delivery_care)) %>%
  filter(continent == "Asia")

head(q2_indicators)
```

```
##   country_code_iso3       country_name continent    sub_region_name hdr_le_2021
## 1               MMR            Myanmar      Asia South-eastern Asia     65.6716
## 2               PSE State of Palestine      Asia       Western Asia          NA
## 3               PSE State of Palestine      Asia       Western Asia          NA
## 4               KAZ         Kazakhstan      Asia       Central Asia     69.3622
## 5               KGZ         Kyrgyzstan      Asia       Central Asia     69.9774
## 6               TJK         Tajikistan      Asia       Central Asia     71.5942
##   under5_mortality_rate delivery_care mcv hepb hib adolescent_mortality
## 1              41.81237          60.2  44   37  37             5.725543
## 2              14.83475          99.7  98   95  95             5.133172
## 3              14.83475          99.7  98   95  95             5.133172
## 4              10.27429          99.9  97   95  95             4.306696
## 5              17.39998          99.8  93   89  88             4.868060
## 6              31.41900          94.8  97   97  97             2.771752
```

```
q3_indicators <- project_data %>%
  filter(!is.na(mcv), !is.na(hepb), !is.na(hib), !is.na(adolescent_mortality)) %>%
  filter(continent == "Asia")

head(q3_indicators)
```

```
##   country_code_iso3       country_name continent    sub_region_name hdr_le_2021
## 1               MMR            Myanmar      Asia South-eastern Asia     65.6716
## 2               PSE State of Palestine      Asia       Western Asia          NA
## 3               PSE State of Palestine      Asia       Western Asia          NA
## 4               KAZ         Kazakhstan      Asia       Central Asia     69.3622
## 5               KGZ         Kyrgyzstan      Asia       Central Asia     69.9774
## 6               TJK         Tajikistan      Asia       Central Asia     71.5942
```

```
##   under5_mortality_rate delivery_care mcv hepb hib adolescent_mortality
## 1              41.81237          60.2  44   37  37             5.725543
## 2              14.83475          99.7  98   95  95             5.133172
## 3              14.83475          99.7  98   95  95             5.133172
## 4              10.27429          99.9  97   95  95             4.306696
## 5              17.39998          99.8  93   89  88             4.868060
## 6              31.41900          94.8  97   97  97             2.771752
```

Research Question 1

Introduction: Understanding global disparities in life expectancy is essential for public health planning, resource allocation, and policy formulation. As one of the most direct indicators of health and well-being, life expectancy mirrors the general health status within a territory/country. By definition from world data bank, life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life. To interpret how Asia's overall health and well-being is, it is imperative to understand the life expectancy of newborns on this continent. Hence, we want to investigate the estimated range of life expectancy in Asia and compare the mean of life expectancy in Asia to the rest of world.

Question: What is the estimated range of mean life expectancy in Asian countries? Is the mean life expectancy in Asian countries consistent with the rest of the world?

Bootstrapping (to investigate the estimate of life expectancy)

(1) Sort the data related to expectancy into two different tables for bootstrapping

```
le_data <- data %>%
  mutate(is_asian = case_when(
    continent == "Asia" ~ "YES",
    continent != "Asia" ~ "NO"
  ))

asian_life_expectancy_data <- le_data %>%
  filter(is_asian == "YES")

non_asian_life_expectancy_data <- le_data %>%
  filter(is_asian == "NO")

asian_le_table <- tibble(asian_life_expectancy_data)
non_asian_le_table <- tibble(non_asian_life_expectancy_data)
```

(2) Create the random samples while making sure that all missing values are removed.

```
set.seed(130)
asian_sample <- asian_le_table %>% sample_n(size = 40, replace = TRUE)
non_asian_sample <- non_asian_le_table %>% sample_n(size = 40, replace = TRUE)
```

(3) Initialize variables for the simulate

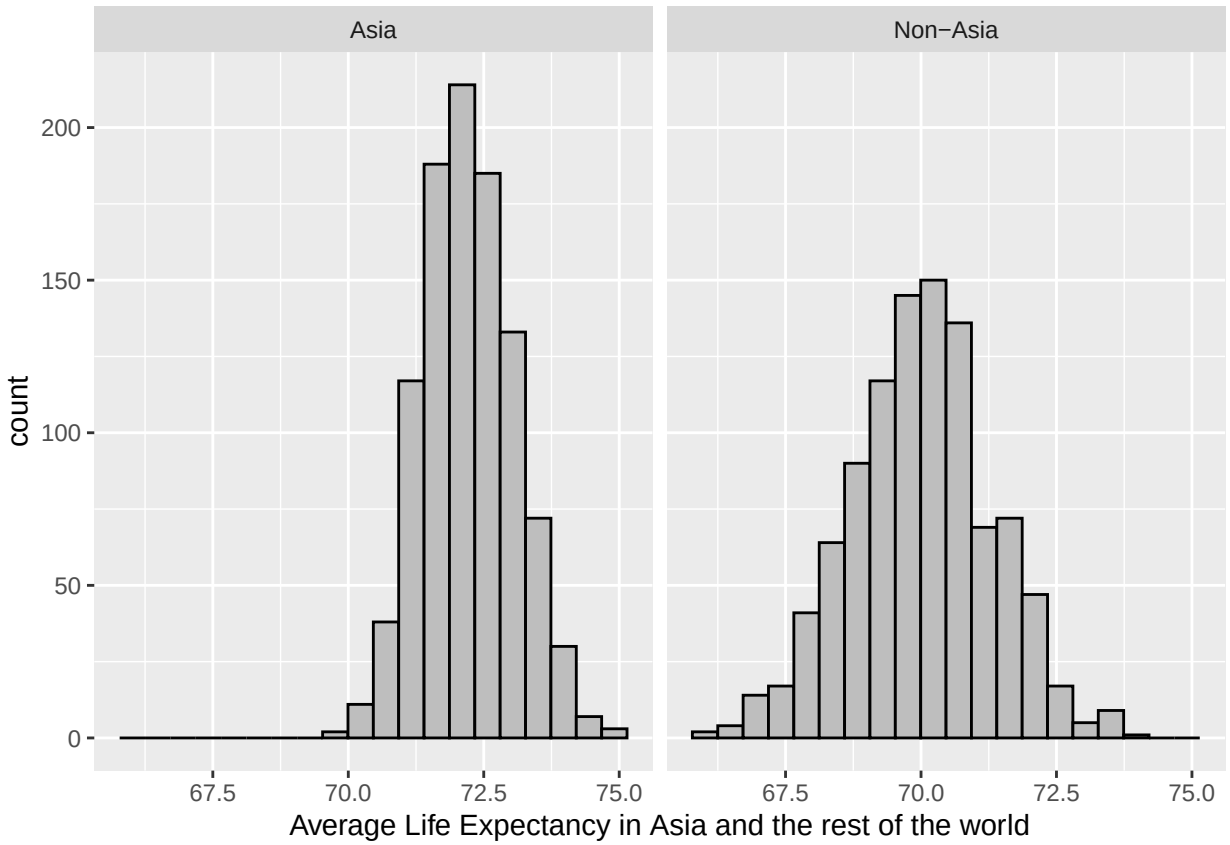
```
num_repetitions <- 1000
asian_simulated_values <- rep(NA, num_repetitions)
non_asian_simulated_values <- rep(NA, num_repetitions)
```

(4) Run the simulation for asian countries

```r
n <- 40
repetitions <- num_repetitions
sim1 <- asian_simulated_values
sim2 <- non_asian_simulated_values
set.seed(130)

for (i in 1:repetitions)
{
  new_sim <- sample(asian_sample$hdr_le_2021 ,size = n, replace=TRUE)
  sim_mean <- mean(new_sim)
  sim1[i] <- sim_mean
}

for (i in 1:repetitions)
{
  new_sim <- sample(non_asian_sample$hdr_le_2021 ,size = n, replace=TRUE)
  sim_mean <- mean(new_sim)
  sim2[i] <- sim_mean
}

sim1 <- tibble(mean = sim1, is_asian = "Asia")
sim2 <- tibble(mean = sim2, is_asian = "Non-Asia")

simulation <- full_join(x = sim1, y = sim2)
```

(5) Plot the simulated means of life expectancy

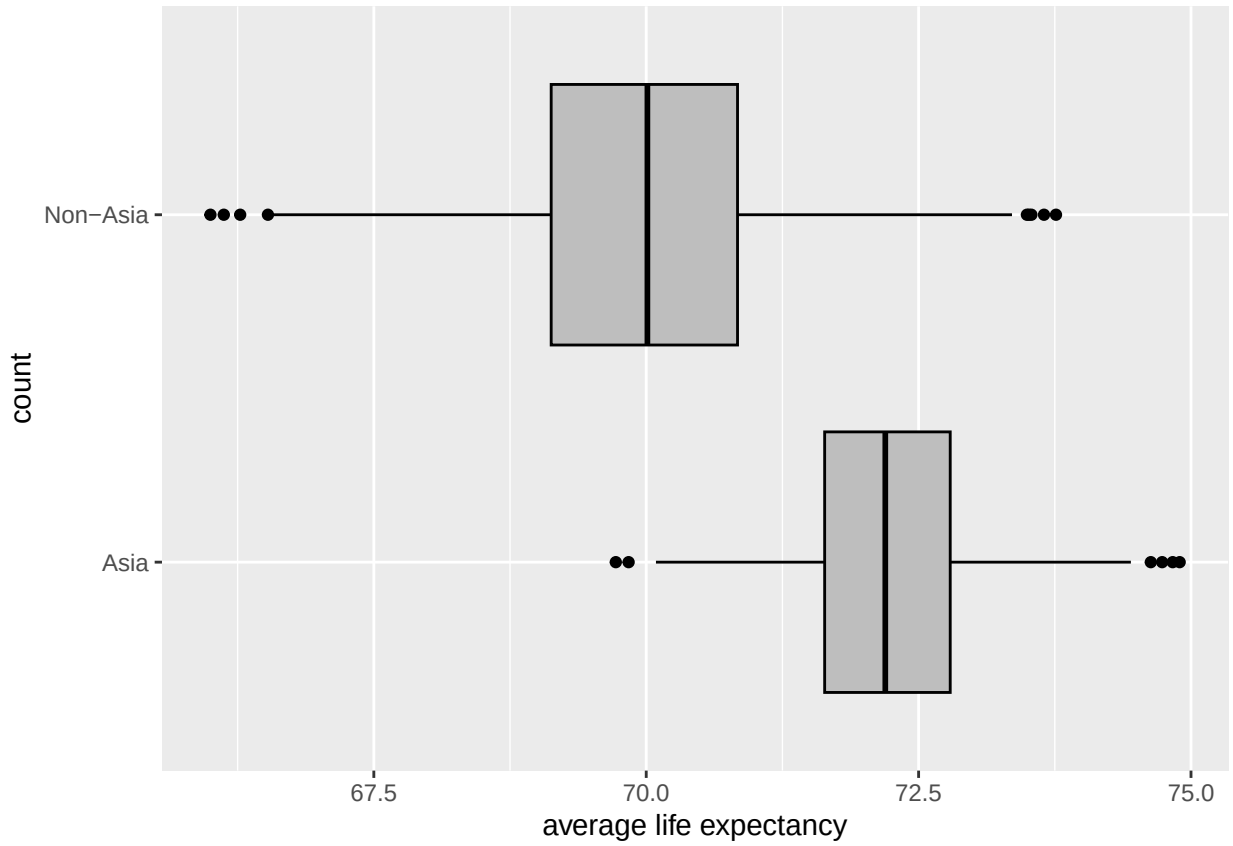
```r
simulation %>% ggplot(aes(x = mean)) +
  geom_histogram(
    color = "black",
    fill = "grey",
    bins = 20
  ) +
  labs(x = "Average Life Expectancy in Asia and the rest of the world") +
  facet_wrap(~is_asian)
```



From Observing and comparing the two histograms, we can find that the sampling distribution in Asia and the rest of the world are both symmetrical and unimodal with very similar spread of distribution.

(6) Make a side-by-side boxplots to better observe the mean, IQR and outliers.

```
simulation %>% ggplot(aes(x = is_asian, y = mean)) +
  geom_boxplot(
    color = "black",
    fill = "grey"
  ) +
  labs(x = "count", y = "average life expectancy") +
  coord_flip()
```



(7) Compute 95% confidence interval for Asian countries

```
asia_lower_bound <- quantile(sim1$mean, p=0.025)
asia_upper_bound <- quantile(sim1$mean, p=0.975)

asia_lower_bound
```

```
##     2.5%
## 70.70934
```

```
asia_upper_bound
```

```
##    97.5%
## 73.87821
```

Base on the result above, we see that the 95% confidence interval for Asian countries are (70.70934, 73.87821)

And also, 95% confidence interval for non-Asian countries

```
non_asian_lower_bound <- quantile(sim2$mean, p=0.025)
non_asian_upper_bound <- quantile(sim2$mean, p=0.975)

non_asian_lower_bound
```

```
##     2.5%
## 67.33909
```

```
non_asian_upper_bound
```

```
##    97.5%
## 72.51791
```

Based on the result above, we see that the 95% confidence interval for non-Asian countries are (67.33909, 72.51791)

It seems like that the average life expectancy in Asia is larger than the rest of the world from the result of boostrapping confidence intervals. So next, we will conduct a two-sample hypothesis testing to find out if that is true.

Hypothesis Testing

(1) Declare the null hypothesis and alternative hypothesis The null hypothesis of this hypothesis test is the average life expectancy in Asia is the same as the rest of the world. The alternative hypothesis of this hypothesis test is the average life expectancy in Asia is larger than the rest of the world. We set the alpha level $\alpha = 0.05$

$$H_0 : \mu_a - \mu_w = 0 \; vs \; H_A : \mu_a - \mu_w > 0$$

where μ_a is the average life expectancy in Asia and μ_w is average life expectancy in other countries in the world.

(2) Compute the test statistic for the hypothesis test.

```
test_stat <- le_data %>%
  group_by(is_asian) %>%
  summarise(means = mean(hdr_le_2021), .groups="drop") %>%
  summarise(value = diff(means)) %>%
  as.numeric()

print(test_stat)
```

```
## [1] 2.408076
```

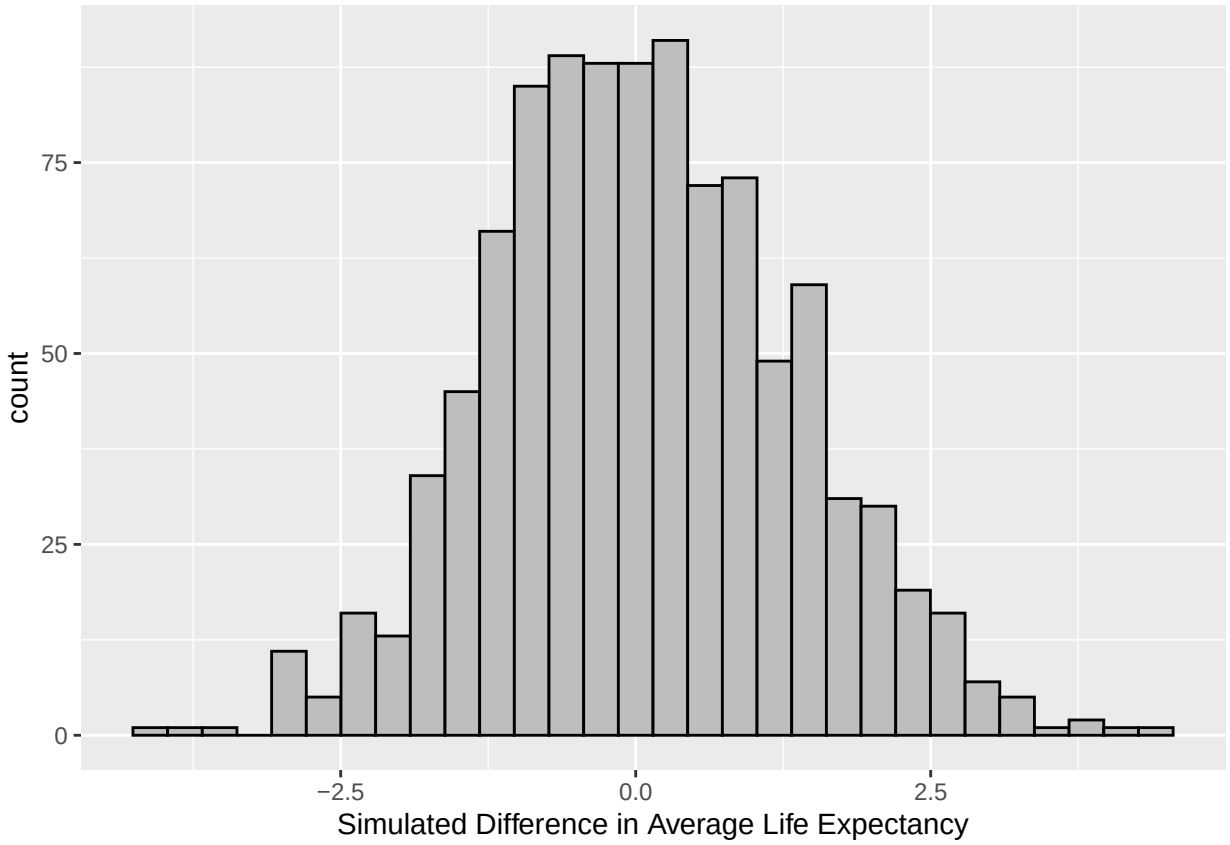
(3) Run the simulation

```
set.seed(130)
n_trials <- 1000
test_stat_simulations <- rep(NA, n_trials)

for(i in 1:n_trials){
  simdata <- le_data %>%
    mutate(is_asian = sample(is_asian), replace = FALSE)
  test_stat_sim <- simdata %>%
    group_by(is_asian) %>%
    summarise(means = mean(hdr_le_2021), .groups="drop") %>%
    summarise(value = diff(means)) %>%
    as.numeric()
  test_stat_simulations[i] <- test_stat_sim
}
```

(4) Plot the sampling distribution of difference in mean of life expectancy between Asia and the rest of the world.

```
tibble(simulated_diff = test_stat_simulations) %>%
  ggplot(aes(x = simulated_diff)) +
  geom_histogram(
    color = "black",
    fill = "grey"
```

```
  ) +
  labs(x = "Simulated Difference in Average Life Expectancy")
```



Finally, we calculate the p-value based on the null hypothesis H_0

```
pvalue <- sum(test_stat_simulations >=  test_stat) / n_trials

print(pvalue)
```

```
## [1] 0.039
```

The p-value of 0.039 is smaller than the alpha level 0.05, so there is moderate/strong evidence against the null hypothesis. As a result, we reject the null hypothesis in favour of the alternative that the average life expectancy in Asia is larger than the average life expectancy of the rest of the world.

Conclusion

In conclusion, the estimated range of life expectancy in Asia is (70.70934, 73.87821). It means that we are 95% confident that the true mean of life expectancy in Asia in 2021 fall between 70.70934 and 73.87821. To be more specific, by 95% confident, it means that 95% of the time, a 95% confidence interval would include the true mean of life expectancy in Asian countries. On the other hand, as a frame of context, the estimated range of life expectancy in non-Asian countries is (67.33909, 72.51791). It means that we are 95% confident that the true mean of life expectancy in non-Asian countries falls between 67.33909 and 72.51791. Through our hypothesis test, the data displayed moderate-to-strong evidence against our null hypothesis that the average life expectancy is the same as the rest of the world. It is safe to conclude that on average, Asian countries' life expectancy is larger than the rest of the world.

With this finding, we are able to proceed our research aimed to find what might have been associated with this

larger life expectancy. Since Asia has a larger life expectancy as the rest of the world, by conducting further research on data of Asian countries, we are able to find valuable information contributing to this large life expectancy and use these information to help the rest of the world in working towards SDG Goal 3 .

Research Question 2

Introduction: Reducing childhood mortality is a key universal health problem. We want to investigate whether we can lower the childhood mortality rate to improve the well-being in Asia by raising the delivery care rate.

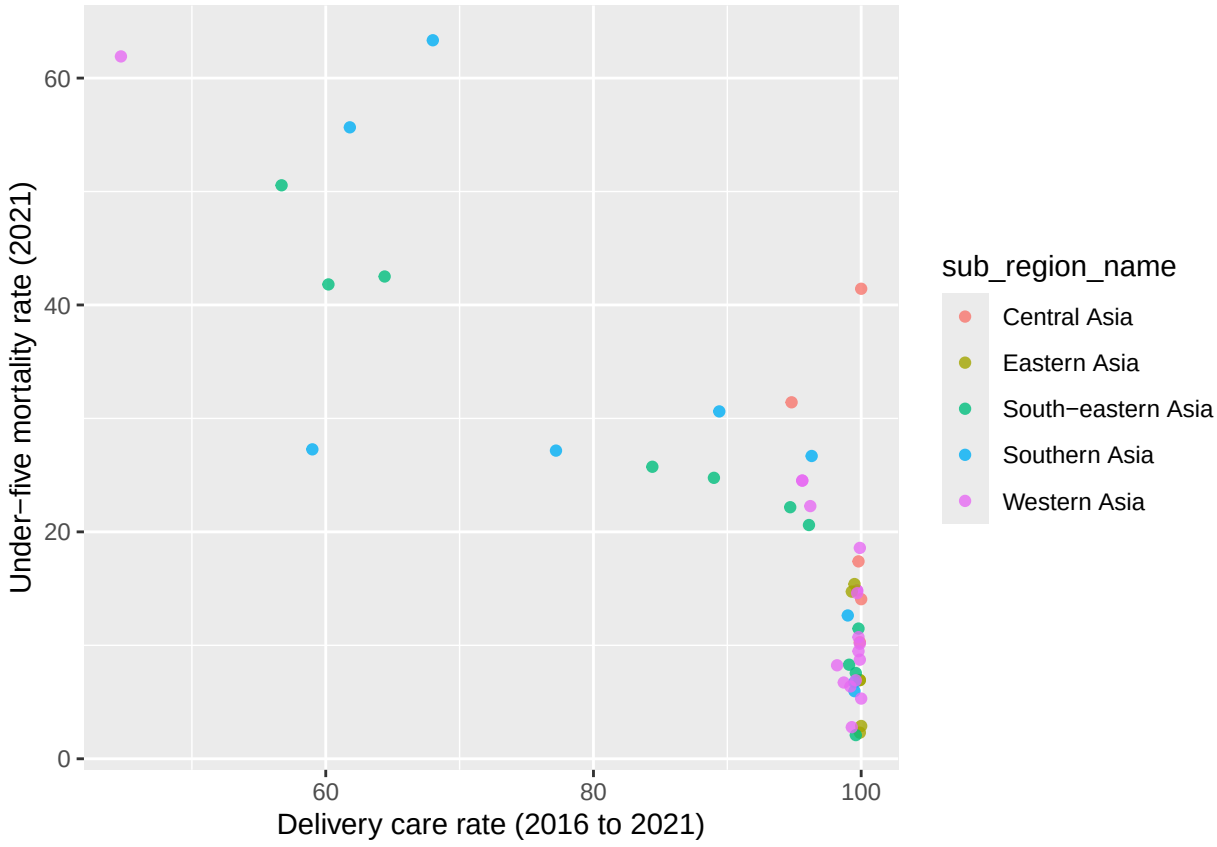
Question: Is there a linear association between the under-five mortality rate and the delivery care rate in Asia? If there is, what is the linear association?

Definitions:(1) Delivery care rate: The proportion of births attended by skilled health personnel as defined as the percentage of live births attended by skilled health personnel (doctor, nurse, midwife).(2) Under-five mortality rate: Probability of dying between birth and exactly 5 years of age, expressed per 1,000 live births.

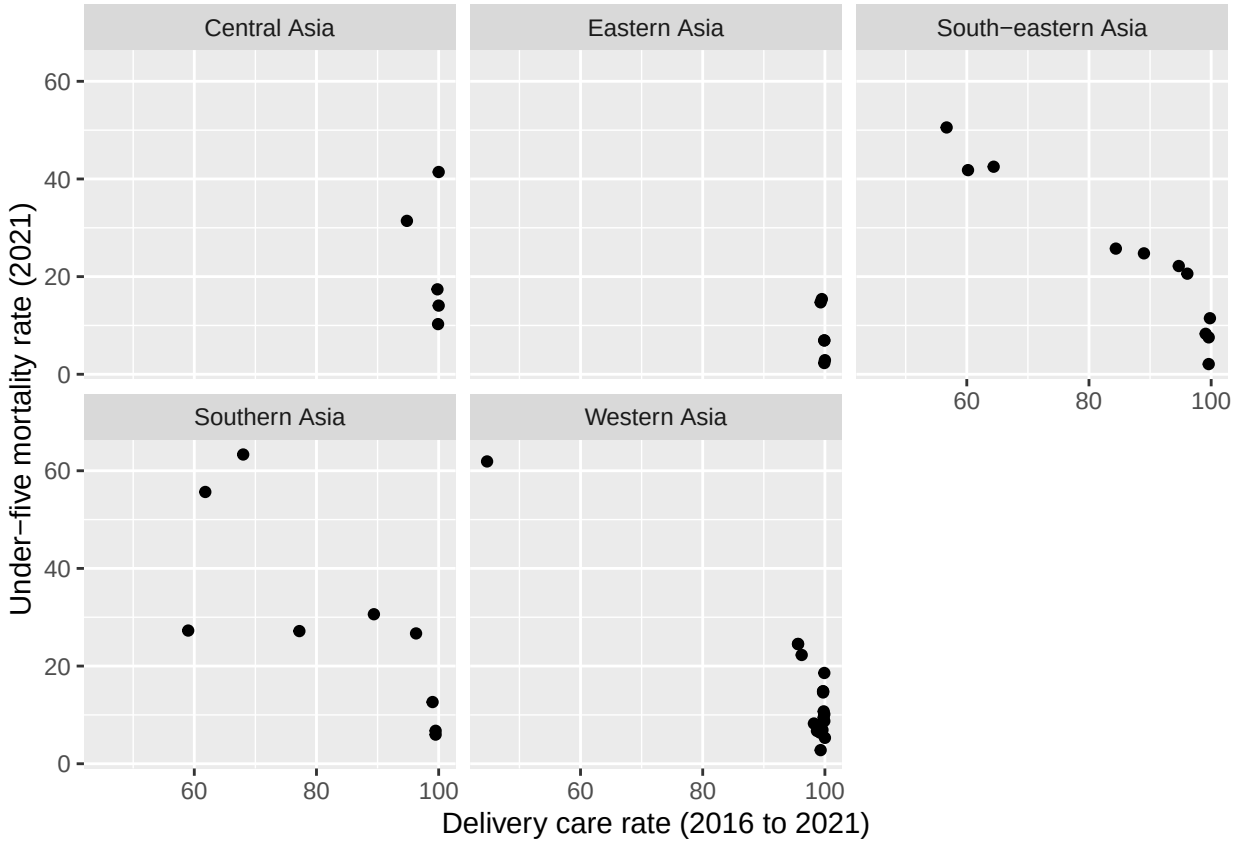
Building model

(1) Visualize the association with scatter plots

```
ggplot(data = q2_indicators, aes(y = under5_mortality_rate,
                                 x = delivery_care, color = sub_region_name)) +
  geom_point(alpha = 0.8) +
  labs(x = "Delivery care rate (2016 to 2021)",
       y = "Under-five mortality rate (2021)")
```



```
ggplot(data = q2_indicators, aes(y = under5_mortality_rate,
                                 x = delivery_care)) +
  geom_point() +
  labs(x = "Delivery care rate (2016 to 2021)",
       y = "Under-five mortality rate (2021)") +
  facet_wrap(~sub_region_name)
```



We can observe that there might be a negative linear relationship between two variables, and the relationships are different among different sub-regions of Asia.

(2) Quantify the association by correlation of different sub-regions

```
q2_indicators %>%
  group_by(sub_region_name) %>%
  summarise(correlation = cor(under5_mortality_rate, delivery_care))
```

```
## # A tibble: 5 x 2
##   sub_region_name    correlation
##   <chr>                    <dbl>
## 1 Central Asia            -0.349
## 2 Eastern Asia            -0.920
## 3 South-eastern Asia      -0.953
## 4 Southern Asia           -0.758
## 5 Western Asia            -0.906
```

We notice that the correlations for different sub-regions vary and Central Asia's magnitude of r is close to zero, which indicates a weak strength of linear association. Hence, we decide to choose Central Asia as our baseline in the model.

(3) Create a multiple linear regression model

(a) Build a model

$$Y_i = \beta_0 + \beta_1 x_1, i + \beta_2 x_2, i + \beta_3 x_3, i + \beta_4 x_4, i + \beta_5 x_5, i + \epsilon_i$$

Y_i: the under-five mortality rate x_1, i: delivery care rate(the proportion of births attended by skilled health personnel) x_2, i: 1 when it refers to Eastern Asia, otherwise 0 x_3, i: 1 when it refers to South-eastern Asia, otherwise 0 x_4, i: 1 when it refers to Southern Asia, otherwise 0 x_5, i: 1 when it refers to Western Asia, otherwise 0 β_0: the under-five mortality rate relates to x = 0 (intercept) β_1: the slope of the baseline β_2: the coefficients of x_2, i β_3: the coefficients of x_3, i β_4: the coefficients of x_4, i β_5: the coefficients of x_5, i

(b) multiple coefficient test(t-test) We need to use hypothesis test to check every coefficients:(i = 1,2,3,4,5)

$$H_0 : \beta_i = 0$$

$$H_0 : \beta_i \neq 0$$

We set α level at 0.05

(c) Check p-value

```
model_1 <- lm(under5_mortality_rate~delivery_care + sub_region_name,
              data =  q2_indicators)
summary(model_1)
```

```
##
## Call:
## lm(formula = under5_mortality_rate ~ delivery_care + sub_region_name,
##     data = q2_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.382  -5.082  -1.127   6.124  20.900
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     113.28169    9.79550  11.565 8.75e-15 ***
## delivery_care                    -0.91369    0.09178  -9.955 9.90e-13 ***
## sub_region_nameEastern Asia     -13.94177    4.98527  -2.797  0.00769 **
## sub_region_nameSouth-eastern Asia -11.48880    4.60031  -2.497  0.01642 *
## sub_region_nameSouthern Asia     -8.71829    4.80960  -1.813  0.07687 .
## sub_region_nameWestern Asia     -10.61112    4.17076  -2.544  0.01463 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.232 on 43 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7219
## F-statistic: 25.91 on 5 and 43 DF,  p-value: 5.746e-12
```

Since the p-value for β_1 are significantly smaller than the alpha value we set, we have strong evidence against the null hypothesis that the slope is zero. Hence, we can conclude that there exists a linear relationship between the under-five mortality rate and the delivery care rate in Central Asian countries. Moreover, we notice that Southern Asia fails the t-test, which means that we fail to reject the Null hypothesis($\beta_4 = 0$), which indicates that the slopes of the regression lines are not parallel for the particular variable Southern Asia, and other variables fit the parallel linear regression lines model well. The multiple R-squared value of 0.75 suggests that, in Asia, 75% of the variance in the under-five mortality rate can be explained by the delivery care rate in the model. The remaining 25% of the variance is unexplained by the model and can be attributed to either variables not included in the model or inherent variability in the data.

We have equations for fitted linear regression lines:

For Central Asia countries:

$$Y_i = 113.28 - 0.91 * delivery\,care\,rate$$

For Eastern Asia:

$$Y_i = 113.28 - 0.91 * delivery\,care\,rate - 13.94$$

For South-eastern Asia:

$$Y_i = 113.28 - 0.91 * delivery\,care\,rate - 11.49$$

For Southern Asia:

$$Fail\,to\,fit\,in\,parallel\,lines\,model$$

For Western Asia:

$$Y_i = 113.28 - 0.91 * delivery\,care\,rate - 10.61$$

(d) Visualize linear regression lines

```
library(broom)
augment(model_1)
```

```
## # A tibble: 49 x 9
##    under5_mortality_rate delivery_care sub_region_name    .fitted .resid   .hat
##                    <dbl>         <dbl> <chr>                <dbl>  <dbl>  <dbl>
##  1                  41.8          60.2 South-eastern Asia   46.8   -4.98 0.172
##  2                  14.8          99.7 Western Asia         11.6    3.26 0.0574
##  3                  14.8          99.7 Western Asia         11.6    3.26 0.0574
##  4                  10.3          99.9 Central Asia         22.0  -11.7  0.200
##  5                  17.4          99.8 Central Asia         22.1   -4.70 0.200
##  6                  31.4          94.8 Central Asia         26.7    4.76 0.202
##  7                  41.4         100   Central Asia         21.9   19.5  0.200
##  8                  14.1         100   Central Asia         21.9   -7.85 0.200
##  9                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## 10                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## # i 39 more rows
## # i 3 more variables: .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

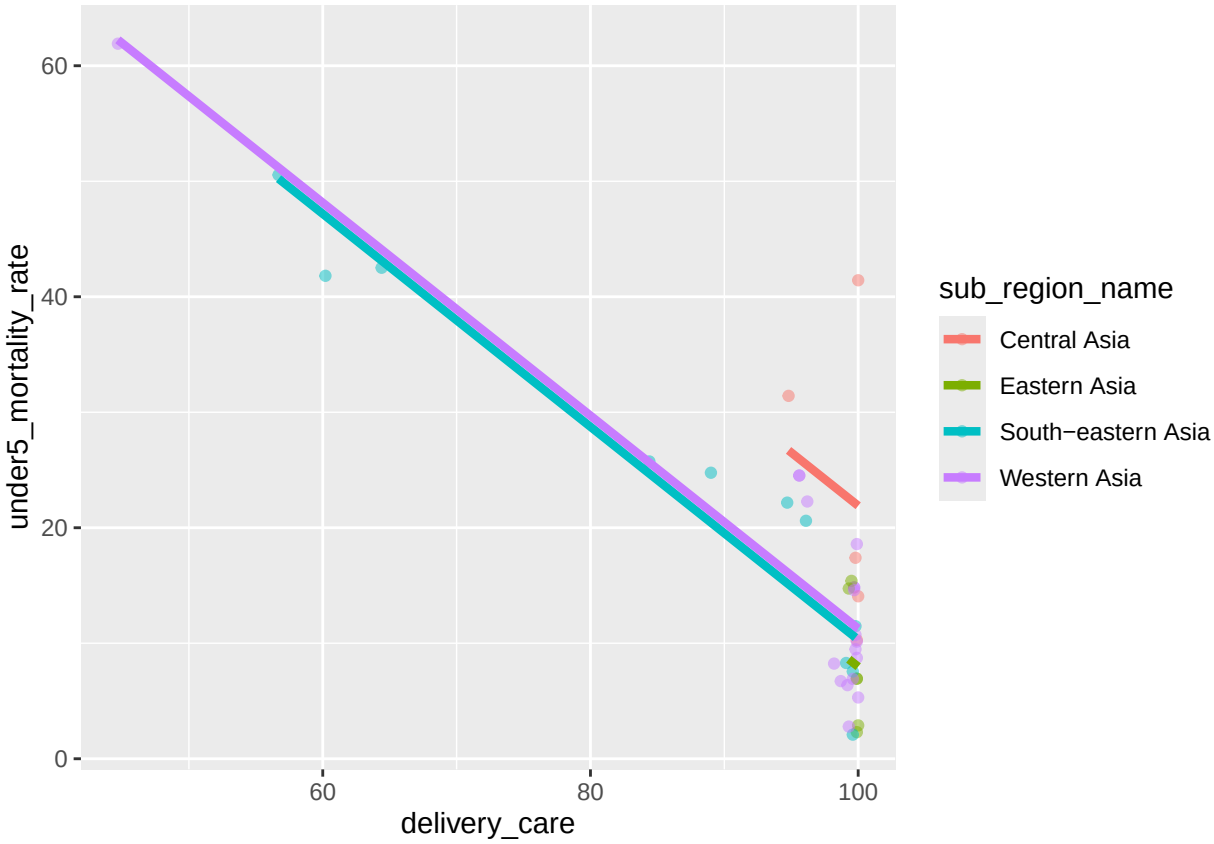
Since Southern Asia case fails the t-test, we remove the Southern Asia Case

```
asia_indicators_modified <- q2_indicators %>%
  filter(sub_region_name != "Southern Asia")
model_1_modified <- lm(under5_mortality_rate~delivery_care + sub_region_name,
              data =  asia_indicators_modified)
library(broom)
augment(model_1_modified)
```

```
## # A tibble: 40 x 9
##    under5_mortality_rate delivery_care sub_region_name    .fitted .resid   .hat
##                    <dbl>         <dbl> <chr>                <dbl>  <dbl>  <dbl>
##  1                  41.8          60.2 South-eastern Asia   47.0   -5.18 0.205
##  2                  14.8          99.7 Western Asia         11.5    3.29 0.0581
##  3                  14.8          99.7 Western Asia         11.5    3.29 0.0581
##  4                  10.3          99.9 Central Asia         22.0  -11.7  0.200
##  5                  17.4          99.8 Central Asia         22.1   -4.69 0.200
##  6                  31.4          94.8 Central Asia         26.7    4.72 0.203
##  7                  41.4         100   Central Asia         21.9   19.5  0.200
##  8                  14.1         100   Central Asia         21.9   -7.84 0.200
##  9                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## 10                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## # i 30 more rows
```

```
## # i 3 more variables: .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
ggplot(data = asia_indicators_modified, aes(y = under5_mortality_rate,
                                            x = delivery_care,
                                            color = sub_region_name)) +
  geom_point(alpha = 0.5) +
  geom_line(data=augment(model_1_modified),
            aes(y=.fitted, colour=sub_region_name), lwd = 1.5)
```




```
  labs(x = "Delivery care rate (2016 to 2021)",
       y = "Under-five mortality rate (2021)") +
  theme_minimal()
```

```
## NULL
```

Conclusion

In conclusion, there is a negative linear relationship between the under-five mortality rate and the delivery care rate among Asian countries in 2021, while the association varies in different sub-regions.

For Central Asian, Southern Asian, South-eastern Asian, and Western Asian countries, increasing the delivery care rate can significantly decrease the under-five mortality rate.

However, in East Asia, the observation of a vertical distribution of Eastern Asia countries' data suggests that while there is a universally high rate of delivery care, there exists significant variability in the under-five mortality rate.This implies that there might be other factors that affect the under-five mortality rate in East Asia. To lower the under-five mortality rate and improve the well-being of children, these countries should shift their focus towards these factors, rather than delivery care rate. According to the research by Madhav Kumar Bhusa and Shankar Prasad Khanal, education of mother, size of child at birth, age of mother at

childbirth, place of residence, and birth interval were the significant and most frequently observed factors of under-five mortality. Nevertheless, specific investigations focusing on the context of East Asian countries remain scarce, presenting an opportunity for future research.

Research Question 3

Questions:Is there a relationship between adolescent mortality of age 10 - 19 and various vaccine (MCV, HepB, Hib) coverage in Asia? Using information based on these two variables in data, what is the linear regression between adolescent mortality and vaccine coverage?

Building model

(1) Linear regression between mcv and adolescent_mortality

(a) Quantify the association by correlation of different sub-regions

```
q3_indicators %>% group_by(continent) %>%
  summarise(cor(x = mcv, y = adolescent_mortality)) # calculate the correlationship between mcv and ado
```

```
## # A tibble: 1 x 2
##   continent `cor(x = mcv, y = adolescent_mortality)`
##   <chr>                                        <dbl>
## 1 Asia                                        -0.393
```

(b) Build a model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Y_i: the adolescent mortality

x_i: mcv coverage rate

β_0: the adolescent mortality rate relates to x = 0

β_1: slope parameter

ϵ_i: random error term for observation

(c) State the null and alternative hypotheses

$$H_0 : \beta_1 = 0 \ vs \ H_A : \beta_1 \neq 0$$

We set α level at 0.05

(d) Check p-value

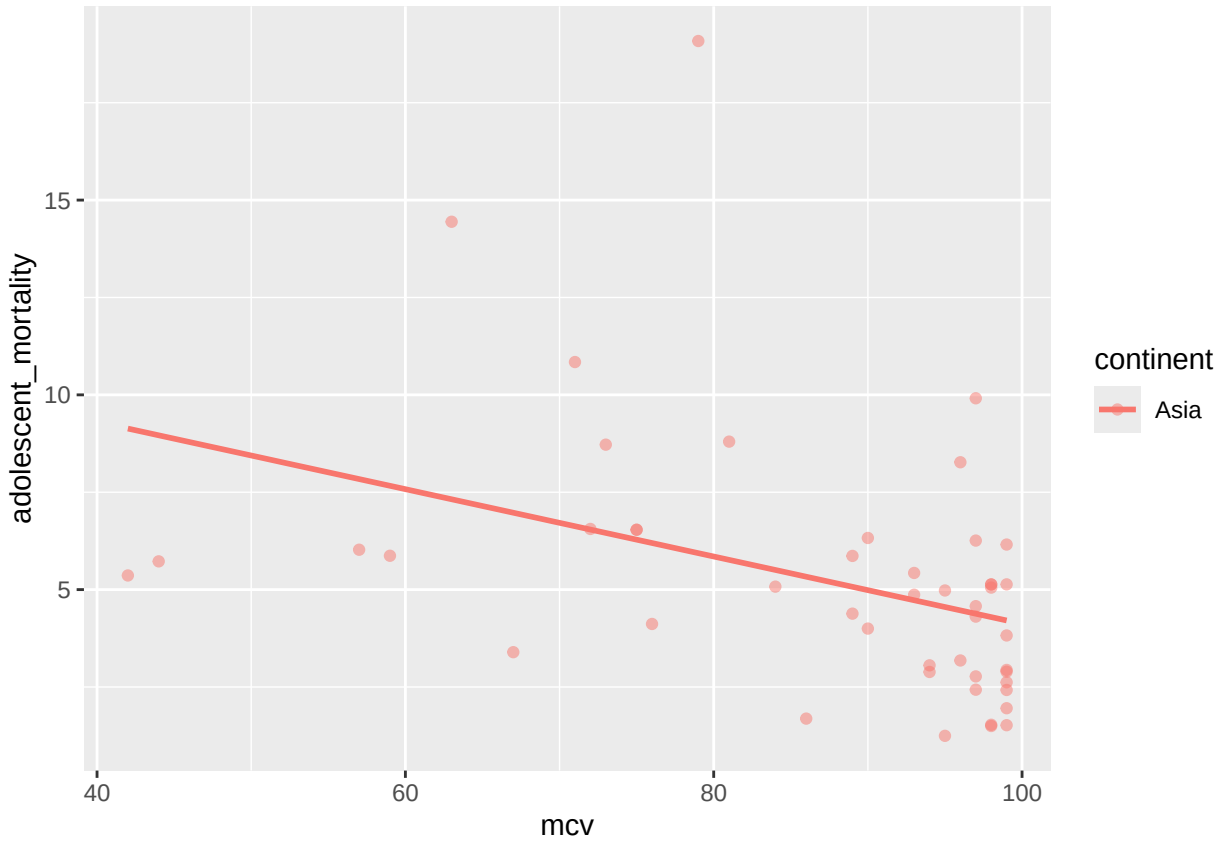
```
mcv_model <- lm(adolescent_mortality ~ mcv, data = q3_indicators) # build a simple linear regression
summary(mcv_model)$coefficients # find the coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.76416483 2.63732872  4.839808 1.501682e-05
## mcv         -0.08642629 0.02984199 -2.896131 5.764076e-03
```

Since the p-value of $\beta_1 = 5.764076\text{e-}03$*is significantly smaller than the alpha value we set, we have strong evidence against the null hypothesis that the slope is zero. Therefore, we can conclude that there is a linear relationship between adolescent mortality and and mcv coverage rate in Asian countries.*

(e) Visualize linear regression lines

```
q3_indicators %>% ggplot(aes(x = mcv, y = adolescent_mortality, color = continent)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) # plot the graph
```



Equation: $\hat{y} = 12.8 - 0.09x_i$

(2) Linear regression between hepb and adolescent_mortality

(a) Quantify the association by correlation of different sub-regions

```
q3_indicators %>% group_by(continent) %>%
  summarise(cor(x = hepb, y = adolescent_mortality)) # calculate the correlationship between hepb and a
```

```
## # A tibble: 1 x 2
##   continent `cor(x = hepb, y = adolescent_mortality)`
##   <chr>                                         <dbl>
## 1 Asia                                         -0.306
```

(b) Build a model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Y_i: the adolescent mortality

x_i: hepb coverage rate

β_0: the adolescent mortality rate relates to x = 0

β_1: slope parameter

ϵ_i: random error term for observation

(c) State the null and alternative hypotheses

$$H_0 : \beta_1 = 0 \ vs \ H_A : \beta_1 \neq 0$$

We set α level at 0.05

(d) Check p-value

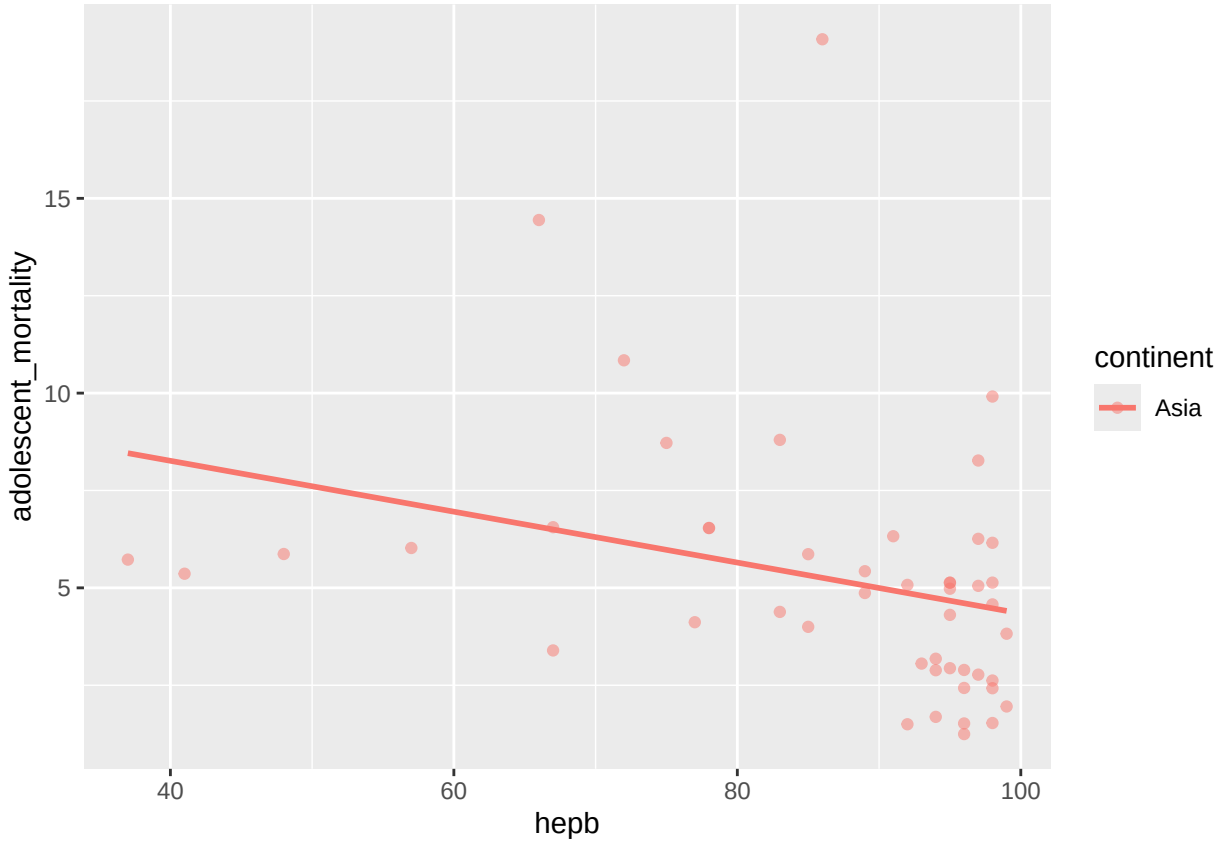
```r
hepb_model <- lm(adolescent_mortality ~ hepb, data = q3_indicators) # build a simple linear regression

summary(hepb_model)$coefficients # find the coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 10.87442072 2.63103068  4.133141 0.0001498363
## hepb        -0.06530903 0.03000246 -2.176789 0.0346655585
```

Since the p-value of $\beta_1 = 0.0346655585$*is significantly smaller than the alpha value we set, we have strong evidence against the null hypothesis that the slope is zero. Therefore, we can conclude that there is a linear relationship between adolescent mortality and and hepb coverage rate in Asian countries.*

(e) Visualize linear regression lines

```r
q3_indicators %>% ggplot(aes(x = hepb, y = adolescent_mortality, color = continent)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) # plot the graph
```



Equation : $\hat{y} = 10.9 - 0.07x_i$

(2) Linear regression between hepb and adolescent_mortality

(a) Quantify the association by correlation of different sub-regions

```r
q3_indicators %>% group_by(continent) %>%
  summarise(cor(x = hib, y = adolescent_mortality)) # calculate the correlationship between hib and ado
```

```
## # A tibble: 1 x 2
##   continent `cor(x = hib, y = adolescent_mortality)`
##   <chr>                                        <dbl>
## 1 Asia                                        -0.336
```

(b) Build a model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Y_i: the adolescent mortality

x_i: hib coverage rate

β_0: the adolescent mortality rate relates to x = 0

β_1: slope parameter

ϵ_i: random error term for observation

(c) State the null and alternative hypotheses

$$H_0 : \beta_1 = 0 \ vs \ H_A : \beta_1 \neq 0$$

We set α level at 0.05

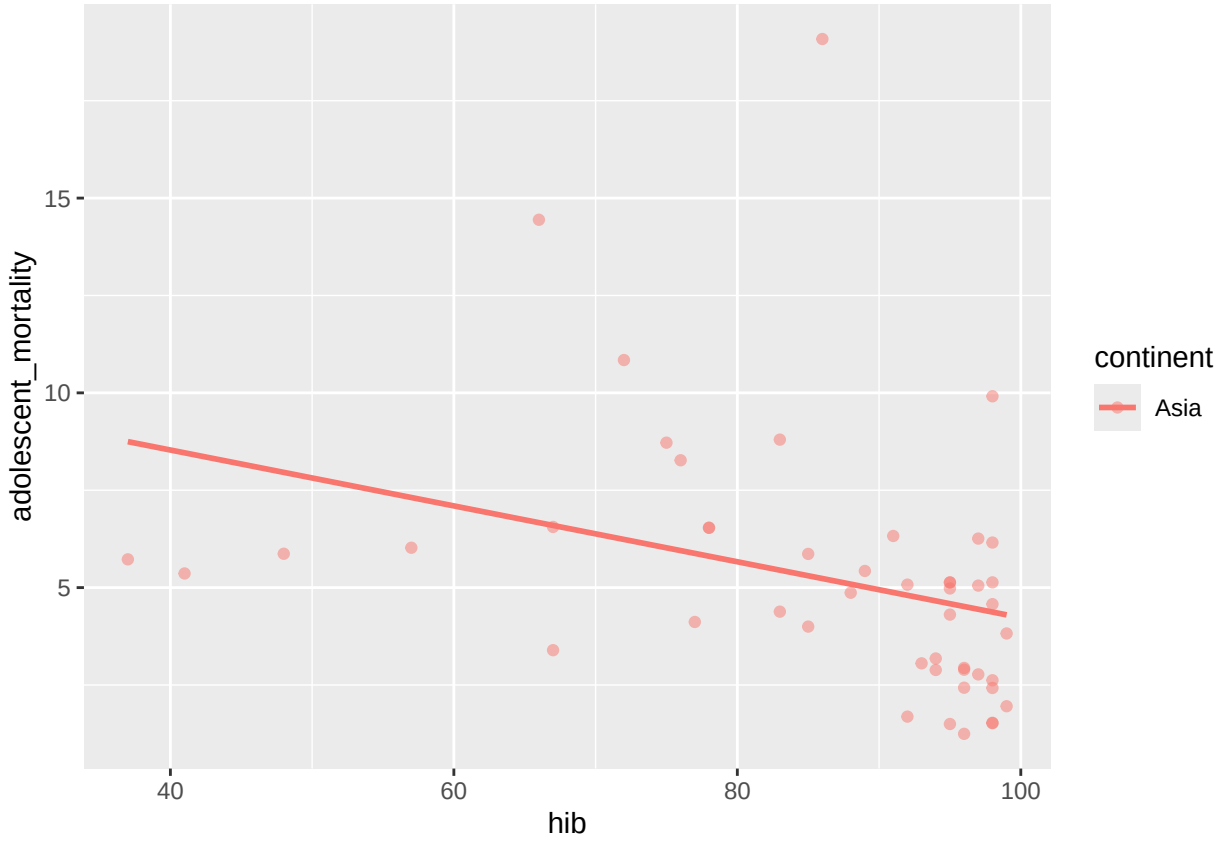
(d) Check p-value

```
hib_model <- lm(adolescent_mortality ~ hib, data = q3_indicators) # build a simple linear regression

summary(hib_model)$coefficients # find the coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 11.40157475 2.58460629  4.411339 6.141893e-05
## hib         -0.07172661 0.02959527 -2.423584 1.935956e-02
```

(e) Visualize linear regression lines

```
q3_indicators %>% ggplot(aes(x = hib, y = adolescent_mortality, color = continent)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) # plot the graph
```



Equation : $\hat{y} = 11.4 - 0.07x_i$

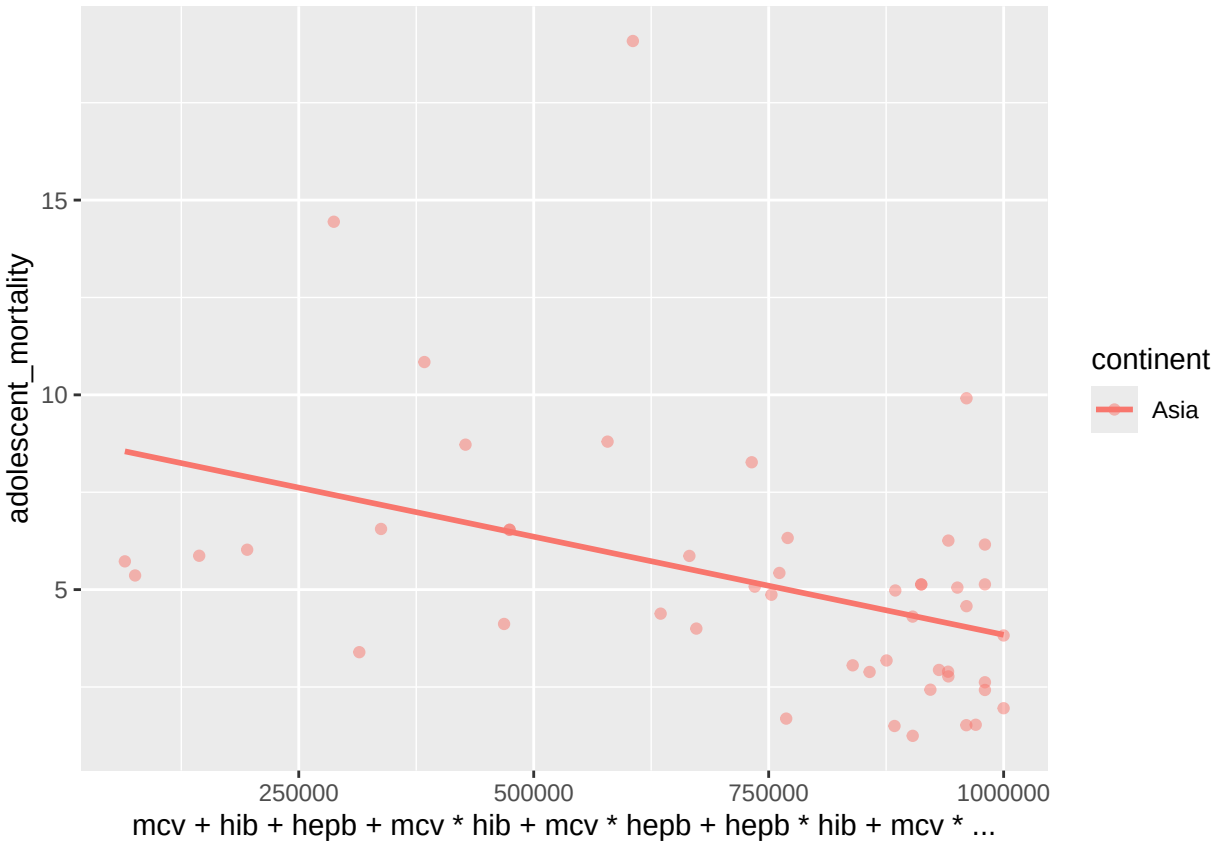
```
library(broom)
augment(model_1)
```

```
## # A tibble: 49 x 9
##    under5_mortality_rate delivery_care sub_region_name    .fitted .resid   .hat
##                    <dbl>         <dbl> <chr>                <dbl>  <dbl>  <dbl>
##  1                  41.8          60.2 South-eastern Asia   46.8   -4.98 0.172
##  2                  14.8          99.7 Western Asia         11.6    3.26 0.0574
##  3                  14.8          99.7 Western Asia         11.6    3.26 0.0574
##  4                  10.3          99.9 Central Asia         22.0  -11.7  0.200
##  5                  17.4          99.8 Central Asia         22.1   -4.70 0.200
##  6                  31.4          94.8 Central Asia         26.7    4.76 0.202
##  7                  41.4         100   Central Asia         21.9   19.5  0.200
##  8                  14.1         100   Central Asia         21.9   -7.85 0.200
##  9                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## 10                   6.93         99.9 Eastern Asia          8.06  -1.13 0.167
## # i 39 more rows
## # i 3 more variables: .sigma <dbl>, .cooksd <dbl>, .std.resid <dbl>
```

```
mcv_model <- lm(adolescent_mortality ~ mcv + hib + hepb + mcv * hib + mcv * hepb + hepb * hib + mcv * h
summary(mcv_model)
```

```
##
## Call:
## lm(formula = adolescent_mortality ~ mcv + hib + hepb + mcv *
##     hib + mcv * hepb + hepb * hib + mcv * hepb * hib, data = q3_indicators)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6634 -1.6794  0.0148  0.9795  9.6084
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.555e+00  3.520e+01   0.101   0.9201
## mcv          -1.094e-01  8.066e-01  -0.136   0.8928
## hib           3.119e+01  1.275e+01   2.446   0.0189 *
## hepb         -3.108e+01  1.298e+01  -2.394   0.0215 *
## mcv:hib      -3.273e-01  1.338e-01  -2.447   0.0189 *
## mcv:hepb      3.265e-01  1.347e-01   2.424   0.0200 *
## hib:hepb      3.738e-03  1.038e-02   0.360   0.7206
## mcv:hib:hepb -2.994e-05  1.077e-04  -0.278   0.7825
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.715 on 40 degrees of freedom
## Multiple R-squared:  0.432,  Adjusted R-squared:  0.3326
## F-statistic: 4.346 on 7 and 40 DF,  p-value: 0.001164
```

```
q3_indicators %>% ggplot(aes(x = mcv + hib + hepb + mcv * hib + mcv * hepb + hepb * hib + mcv * hepb * h
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) # plot the graph
```



Equation:

$\hat{y} = (3.555) + (-0.109)mcv_i + (31.19)hib_i + (-31.08)hepb_i + (-0.327)mcv_i * hib_i + (0.327)mcv_i * hepb_i + (0.004)hib_i * hepb_i + (-0.000$

Conclusion

In conclusion, the inverse association between mortality and vaccination rates for MCV (measles vaccine), HepB (hepatitis B vaccine) and Hib (Haemophilus influenzae type b vaccine) among adolescents aged 10-19 years in Asia is a significant finding. It also reflects Asian countries' progress toward SDG Goal 3 about Good Health and Well-Being.

This correlation suggests that higher vaccine coverage is associated with lower adolescent mortality, suggesting the potential effectiveness of vaccination programs in reducing adolescent mortality in Asia. These findings underscore the importance of prioritizing and scaling up vaccination efforts to ensure broad coverage and access to these essential vaccines.

Furthermore, this correlation highlights the critical role of immunization in preventing vaccine-preventable diseases and their associated complications, which can help reduce mortality in adolescents. It emphasized the need for continued investment in vaccination campaigns, health infrastructure and education initiatives to promote vaccine acceptance and uptake among Asian communities.

Overall, these findings highlight the importance of maintaining and strengthening vaccination efforts as a cornerstone of public health strategies aimed at improving health outcomes and reducing mortality among adolescents in Asia.

References: (1) Databank. (n.d.). https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SP.DYN.LE00.IN#:~:text=Life%20expectancy%20at%20birth%20indicates,the%20same%20throughout%2
(2) Bhusal MK, Khanal SP. A Systematic Review of Factors Associated with Under-Five Child Mortality. Biomed Res Int. 2022 Dec 5;2022:1181409. doi: 10.1155/2022/1181409. PMID: 36518629; PMCID: PMC9744612.