QALL401: Data Analysis for Researchers

Course Contents

1. 2022.12.07: Introduction: About the course [lead by TK]
   - An introduction to open and public data, and data science
2. 2022-12-14: Exploratory Data Analysis (EDA) 1 [lead by hs]
   - R Basics with RStudio and/or RStudio.cloud; Toy Data
3. 2022-12-21: Exploratory Data Analysis (EDA) 2 [lead by hs]
   - R Markdown, `tidyverse` I: `dplyr`; `gapminder`
4. 2023-01-11: Exploratory Data Analysis (EDA) 3 [lead by hs]
   - `tidyverse`II: `readr`, `ggplot2`; Public Data, WDI, WIR, etc
5. 2023-01-18: Exploratory Data Analysis (EDA) 4 [lead by hs]
   - `tidyverse` III: `tidyr`, etc.; WDI, WIR, etc
6. **2023-01-25: Exploratory Data Analysis (EDA) 5 [lead by hs]**
   - `tidyverse` IV; WDI, WIR, etc
7. 2023-02-01: Introduction to PPDAC
   - Problem-Plan-Data-Analysis-Conclusion Cycle: [lead by TK]
8. 2023-02-08: Model building I [lead by TK]
   - Collecting and visualizing data and Introduction to WDI
     (World Development Indicators by World Bank)
9. 2023-02-15: Model building II [lead by TK]
   - Analyzing data and communications
10. 2023-02-22: Project Presentation

1 Exploratory Data Analysis (EDA) I

2 Exploratory Data Analysis II

3 Exploratory Data Analysis III

4 Exploratory Data Analysis (EDA) IV

5 Exploratory Data Analysis (EDA) V

Setup

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(WDI)
library(readxl)
library(broom)
```

- broom: https://cran.r-project.org/web/packages/broom/index.html
- Introduction to broom: https://cran.r-project.org/web/packages/broom/vignettes/broom.html

5.1 Modeling

5.1.1 What is modeling in EDA

Model is a simple summary of data

Goal: A simple low-dimensional summary of a dataset. Ideally, the model will capture true "signals" (i.e. patterns generated by the phenomenon of interest), and ignore "noise" (i.e. random variation that you're not interested in).

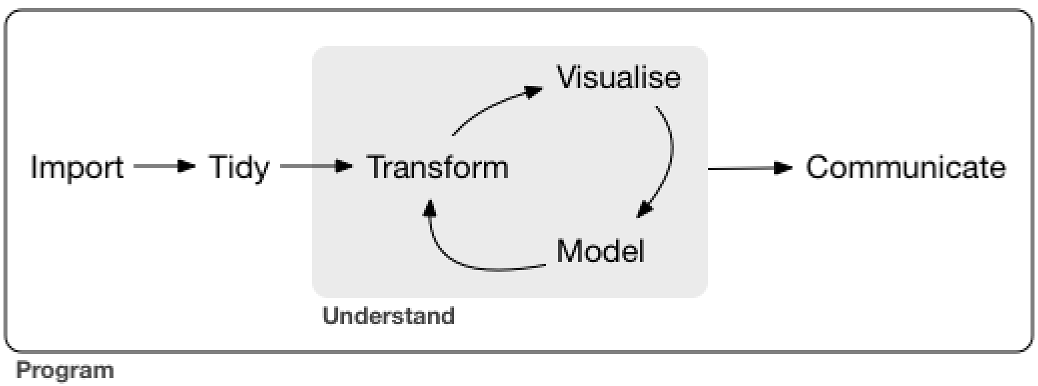
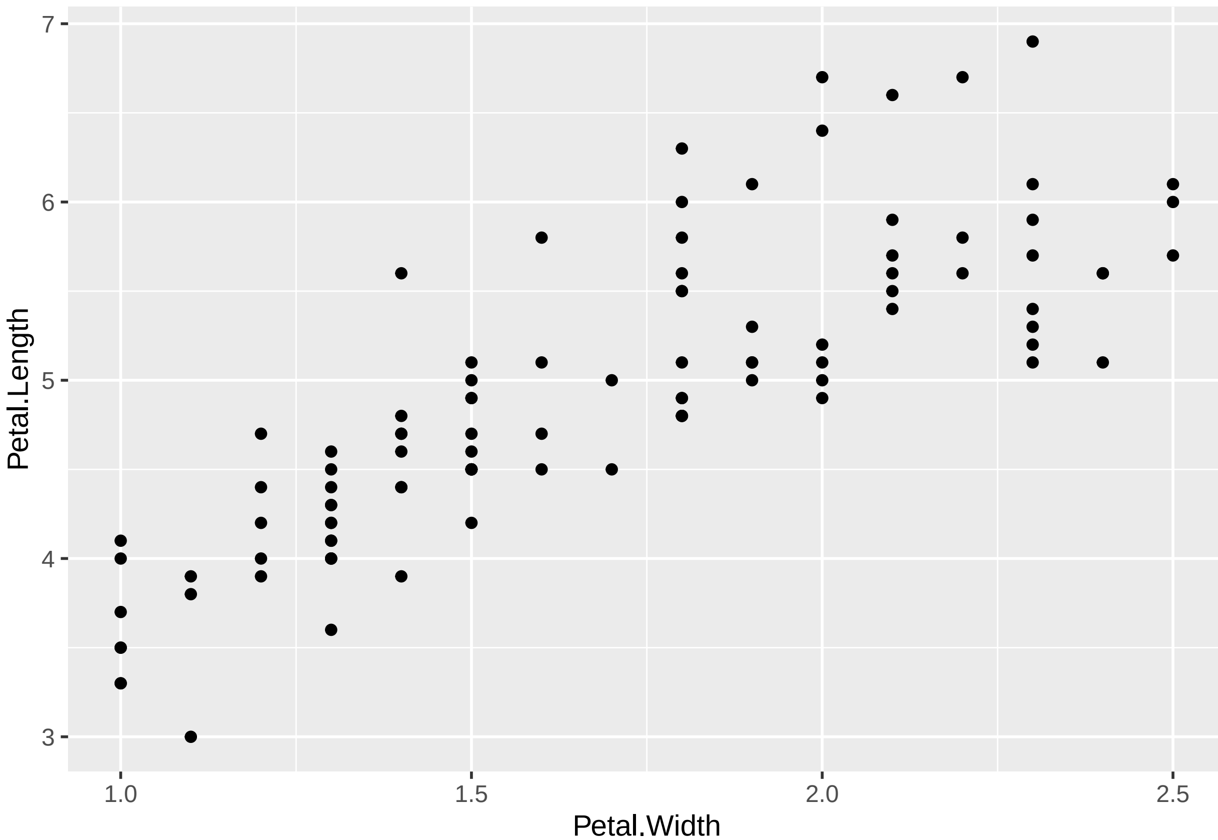


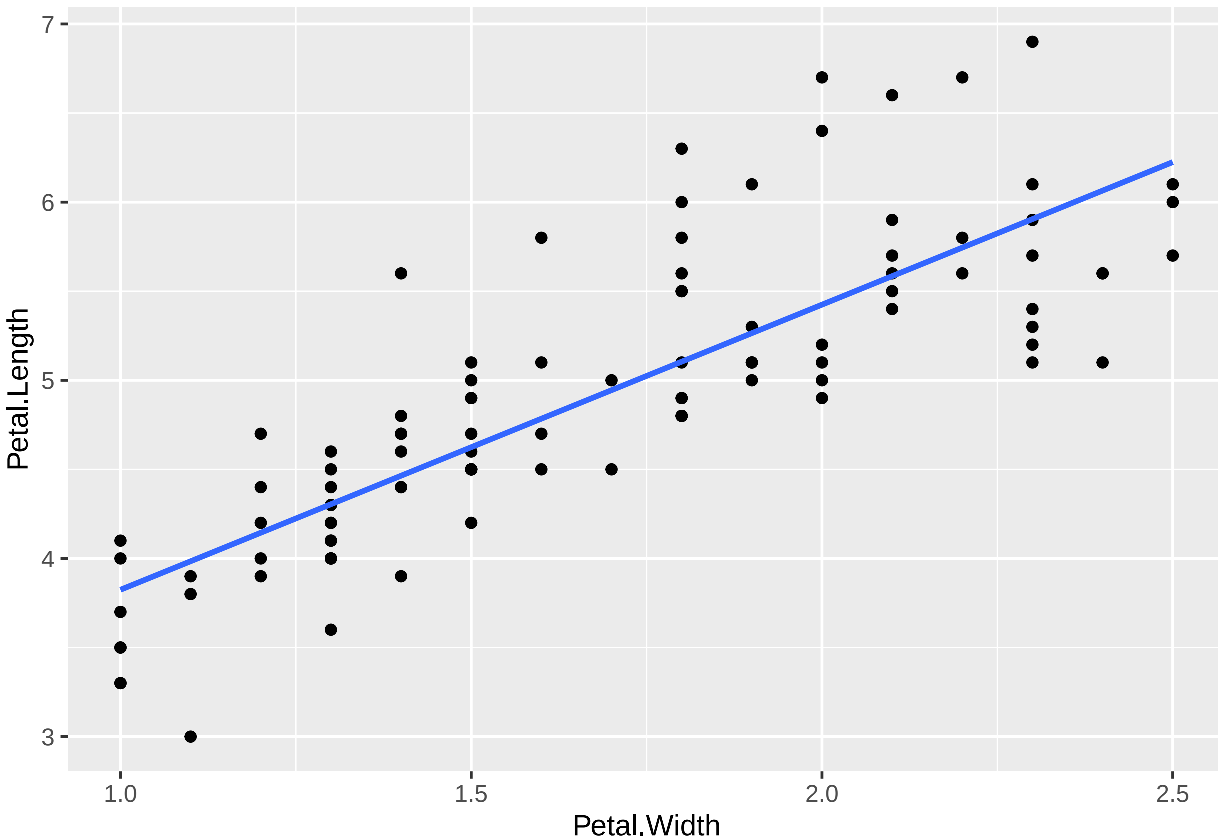
Figure 1: image

```r
df0 <- as_tibble(iris) %>% filter(Species != "setosa")
```

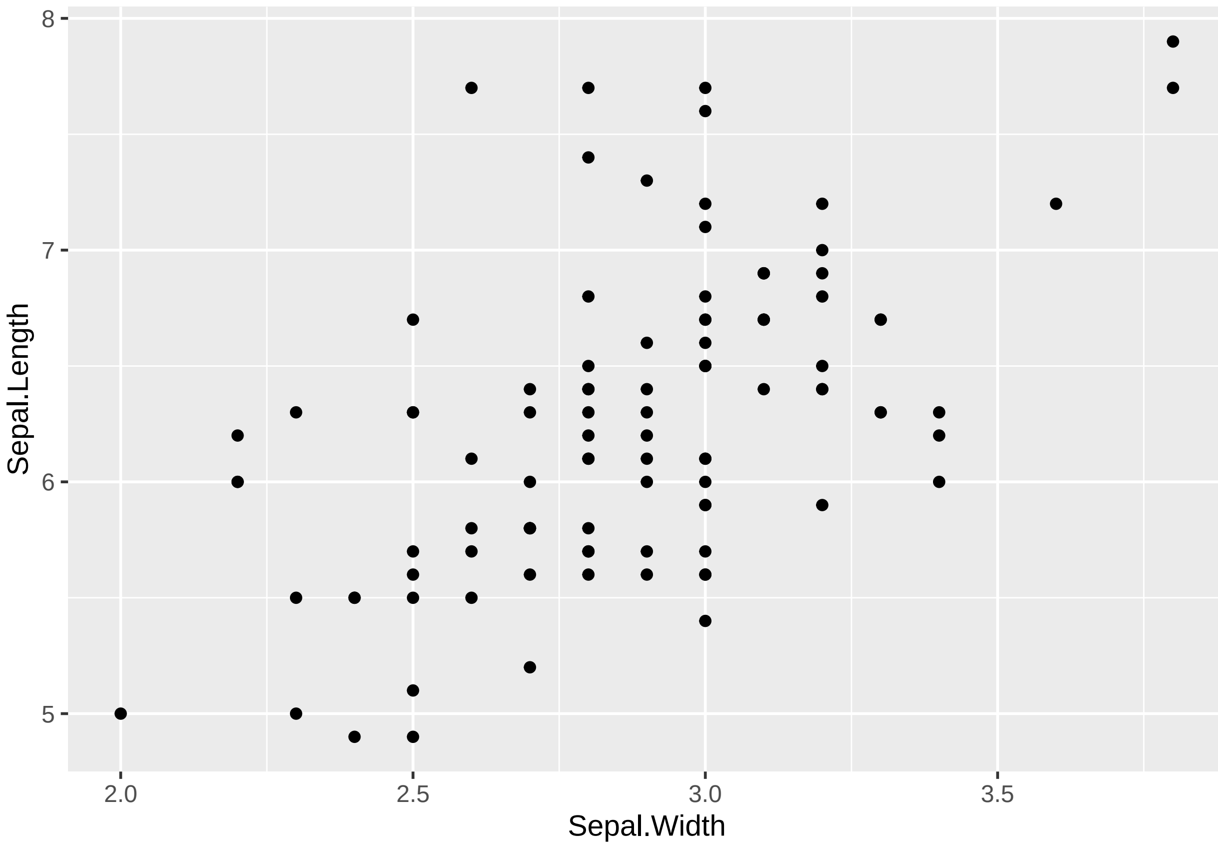
```r
df0 %>% ggplot(aes(Petal.Width, Petal.Length)) + geom_point()
```



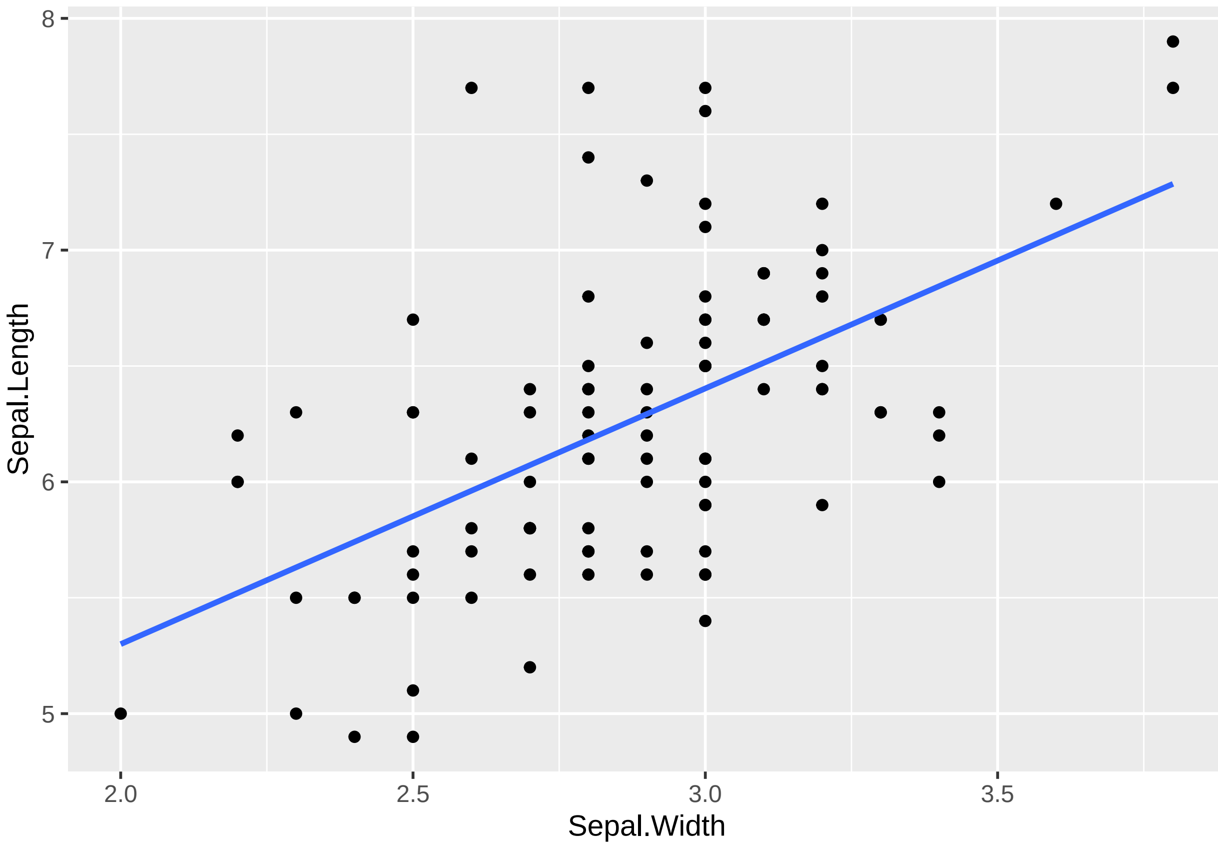
```
df0 %>% ggplot(aes(Petal.Width, Petal.Length)) + geom_point() + geom_smooth(method="lm",formula=y~x, se
```



```
df0 %>% ggplot(aes(Sepal.Width, Sepal.Length)) + geom_point()
```




```
df0 %>% ggplot(aes(Sepal.Width, Sepal.Length)) + geom_point() + geom_smooth(method="lm",formula=y~x, se
```

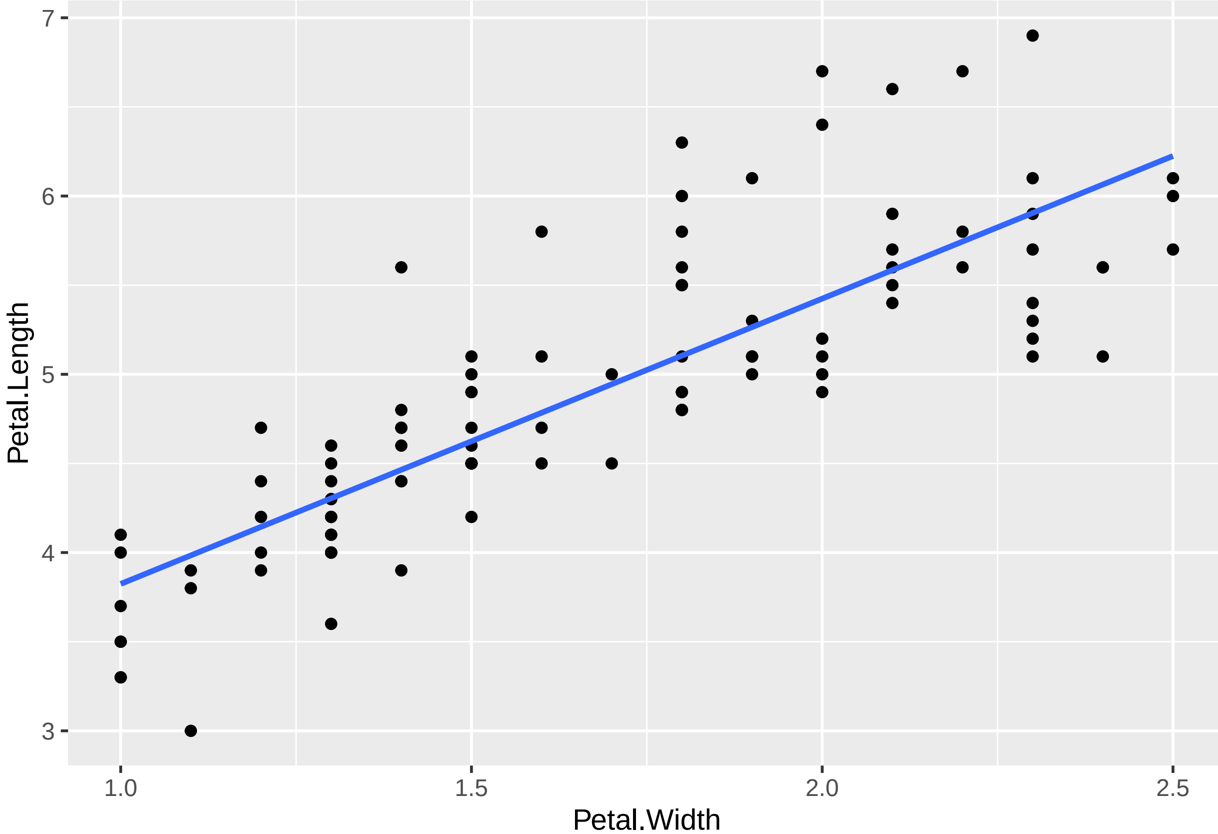


5.1.2 Linear Model: Petal.Length ~ Petal.Width

```
df0 %>% lm(Petal.Length ~ Petal.Width, .)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = .)
##
## Coefficients:
## (Intercept)  Petal.Width
##       2.224        1.600
```

5.1.3 Formula: Petal.Length $= 2.224 + 1.600 \cdot$ Petal.Width

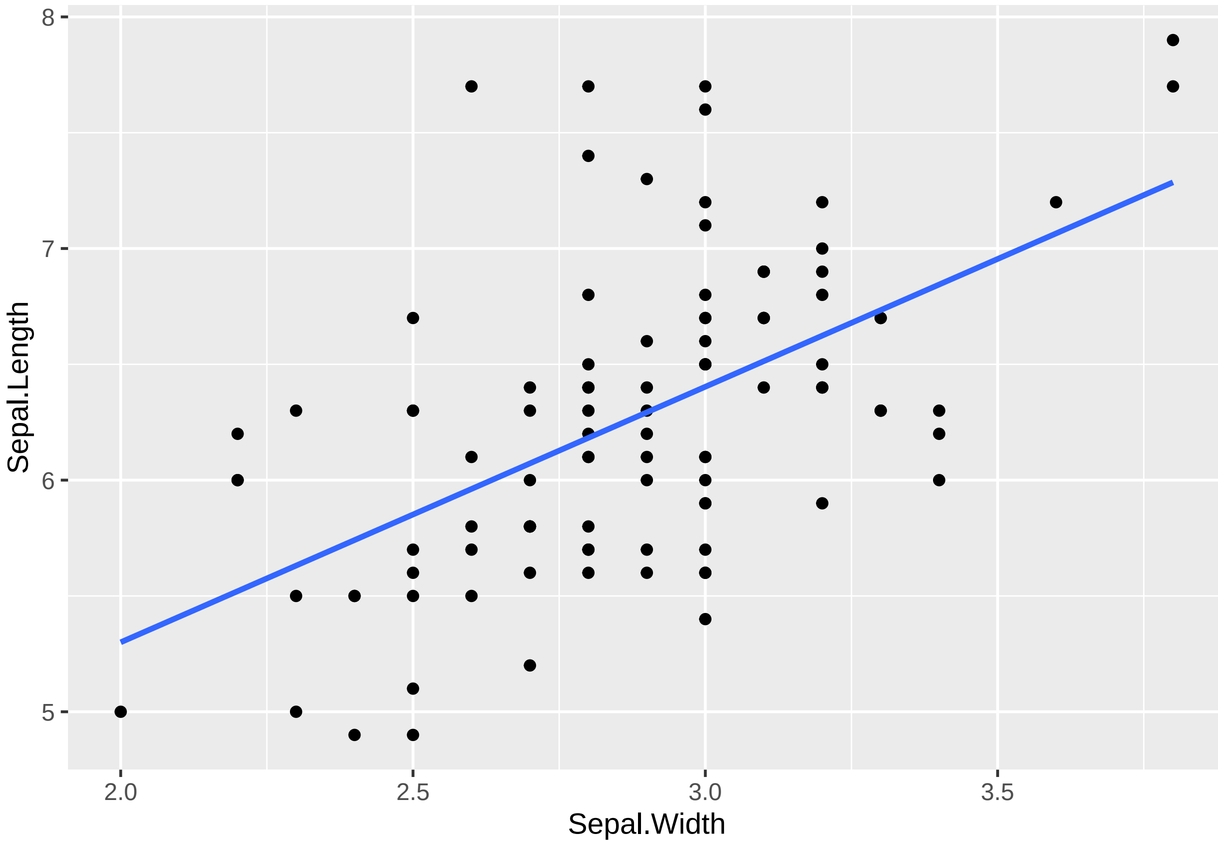


5.1.4 Linear Model: Sepal.Length ~ Sepal.Width

```
df0 %>% lm(Sepal.Length ~ Sepal.Width, .)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width, data = .)
##
## Coefficients:
## (Intercept)  Sepal.Width
##       3.093        1.103
```

5.1.5 Formula: Sepal.Length $= 3.093 + 1.103 \cdot$ Sepal.Width



5.1.6 Petal.Length ~ Petal.Width: R squared = 0.6779 - 68%

```
df0 %>% lm(Petal.Length ~ Petal.Width, .) %>% summary()
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9842 -0.3043 -0.1043  0.2407  1.2755
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.2240     0.1926   11.55   <2e-16 ***
## Petal.Width   1.6003     0.1114   14.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4709 on 98 degrees of freedom
## Multiple R-squared:  0.6779, Adjusted R-squared:  0.6746
## F-statistic: 206.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

5.1.7 Sepal.Length ~ Sepal.Width: R squared = 0.3068 - 31%

```
df0 %>% lm(Sepal.Length ~ Sepal.Width, .) %>% summary()
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0032 -0.3877 -0.0774  0.3200  1.7381
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0934     0.4844   6.387 5.70e-09 ***
## Sepal.Width   1.1033     0.1675   6.585 2.27e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5547 on 98 degrees of freedom
## Multiple R-squared:  0.3068, Adjusted R-squared:  0.2997
## F-statistic: 43.36 on 1 and 98 DF,  p-value: 2.27e-09
```

5.1.8 Linear Model Basics: y ~ x

`lm(y~x, data)`

`data %>% lm(y~x, .)`

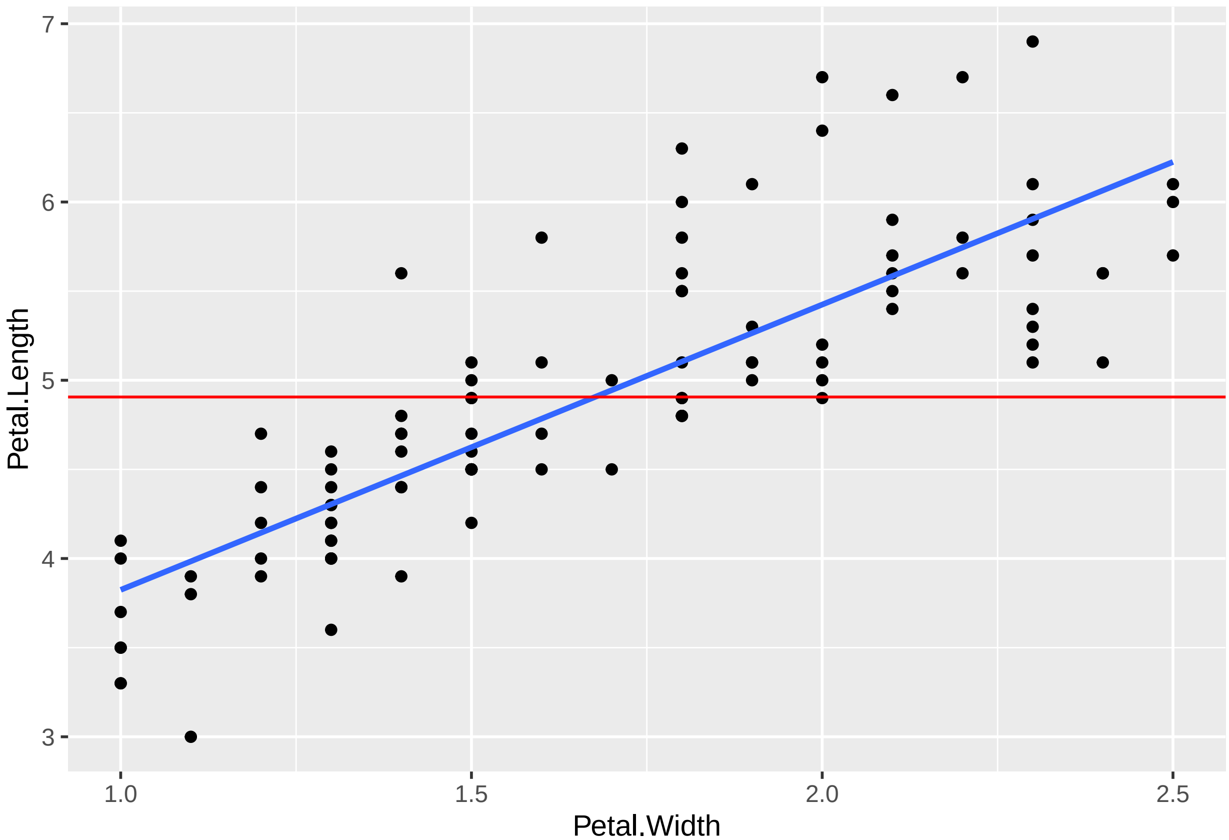
y-intercept, and slope: rate of increase or decrease

`summary(lm(y~x, data))`

`data %>% lm(y~x, .) %>% summary()`

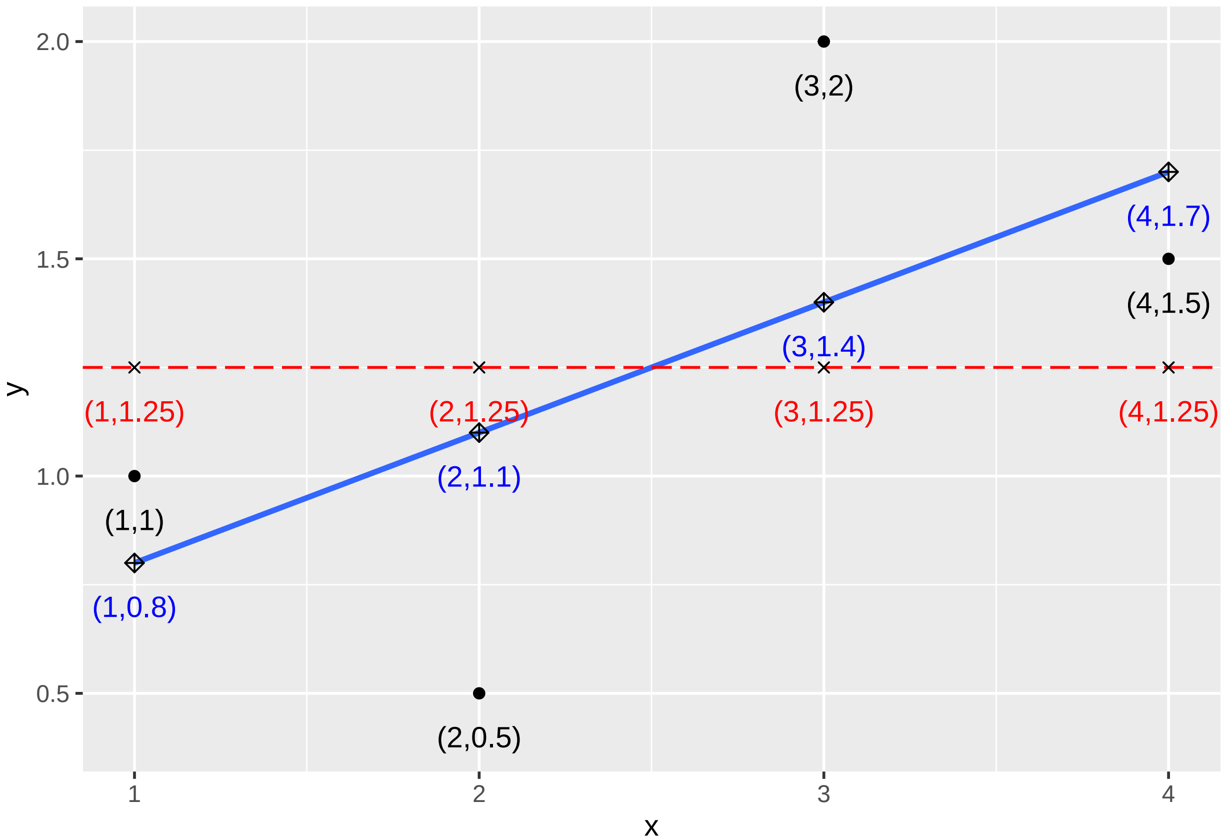
(Multiple) R Squared: a value between 0 and 1, strength of the model

---



```
df1 <- data.frame(x = c(1,2,3,4), y = c(1,0.5,2, 1.5))
ybar <- mean(df1$y)
mod1 <- lm(y~x, df1)
augment(mod1) %>% ggplot() + geom_point(aes(x,y)) + geom_smooth(aes(x,y), formula = y~x, method = "lm",
```

- (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4): Data points
- $\bar{y}$: mean of y $= (y_1 + y_2 + y_3 + y_4)/4$.
- $\hat{y}_i$: prediction at x_i,
  - $(x_1, \hat{y}_1)$, $(x_2, \hat{y}_2)$, $(x_3, \hat{y}_3)$, $(x_4, \hat{y}_4)$ are on the regression line.
- $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$, $y_2 - \hat{y}_2$, $y_2 - \hat{y}_2$ are called residues.



5.1.9 R Squared

$$SS_{tot} = (1-1.25)^2 + (0.5-1.25)^2 + (2-1.25)^2 + (1.5-1.25)^2 = 1.25$$

$$SS_{res} = (1-0.8)^2 + (0.5-1.1)^2 + (2-1.4)^2 + (1.5-1.7)^2 = 0.8$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{0.8}{1.25} = 0.36.$$

```
summary(mod1)$r.squared
```

```
## [1] 0.36
```

```
mod1 %>% glance() %>% pull(r.squared)
```

```
## [1] 0.36
```

```
mod1 %>% glance() %>% select(`R Squared` = r.squared)
```

```
## # A tibble: 1 x 1
##   `R Squared`
##         <dbl>
## 1        0.36
```

```
mod1 %>% summary() %>% glimpse()
```

```
## List of 11
##  $ call         : language lm(formula = y ~ x, data = df1)
##  $ terms        :Classes 'terms', 'formula'  language y ~ x
```

```
##   .. ..- attr(*, "variables")= language list(y, x)
##   .. ..- attr(*, "factors")= int [1:2, 1] 0 1
##   .. .. ..- attr(*, "dimnames")=List of 2
##   .. ..- attr(*, "term.labels")= chr "x"
##   .. ..- attr(*, "order")= int 1
##   .. ..- attr(*, "intercept")= int 1
##   .. ..- attr(*, "response")= int 1
##   .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
##   .. ..- attr(*, "predvars")= language list(y, x)
##   .. ..- attr(*, "dataClasses")= Named chr [1:2] "numeric" "numeric"
##   .. .. ..- attr(*, "names")= chr [1:2] "y" "x"
##  $ residuals    : Named num [1:4] 0.2 -0.6 0.6 -0.2
##   ..- attr(*, "names")= chr [1:4] "1" "2" "3" "4"
##  $ coefficients : num [1:2, 1:4] 0.5 0.3 0.775 0.283 0.645 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "x"
##   .. ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
##  $ aliased      : Named logi [1:2] FALSE FALSE
##   ..- attr(*, "names")= chr [1:2] "(Intercept)" "x"
##  $ sigma        : num 0.632
##  $ df           : int [1:3] 2 2 2
##  $ r.squared    : num 0.36
##  $ adj.r.squared: num 0.04
##  $ fstatistic   : Named num [1:3] 1.12 1 2
##   ..- attr(*, "names")= chr [1:3] "value" "numdf" "dendf"
##  $ cov.unscaled : num [1:2, 1:2] 1.5 -0.5 -0.5 0.2
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "(Intercept)" "x"
##   .. ..$ : chr [1:2] "(Intercept)" "x"
##  - attr(*, "class")= chr "summary.lm"
```

5.1.10 Useful Mathematical Formula

- Let $x = c(x_1, x_2, \ldots, x_n)$ be the independent variable, i.e., Sepal.L
- Let $y = c(y_1, y_2, \ldots, y_n)$ be the dependent variable, i.e., Sepal.W
- Let $\text{pred} = c(\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$ be the predicted values by linear regression.

$$\text{slope of the regression line} = \frac{cov(x,y)}{var(x)} = \frac{cor(x,y)\sqrt{var(y)}}{\sqrt{var(x)}}$$

$$\text{total sum of squares} = SS_{tot} = \sum_i (y_i - mean(y))^2$$

$$\text{residual sum of squares} = SS_{res} = \sum_i (y_i - \text{pred}_i)^2 = \sum_i (y_i - \hat{y}_i)^2$$

$$\text{R squared} = R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = cor(x,y)^2$$

5.1.11 Adjusted R Squared

$$\text{Adjusted } R^2 = 1 - \frac{(1-R^2)(n-1)}{n-k-1}$$

n: number of observations, the number of rows

k: number of variables used for prediction

```
df0 %>% select(1:4) %>% cor()
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.5538548    0.8284787   0.5937094
## Sepal.Width     0.5538548   1.0000000    0.5198023   0.5662025
## Petal.Length    0.8284787   0.5198023    1.0000000   0.8233476
## Petal.Width     0.5937094   0.5662025    0.8233476   1.0000000
```

```
cormat <- df0 %>% select(1:4) %>% cor()
cormat*cormat
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.3067552    0.6863769   0.3524909
## Sepal.Width     0.3067552   1.0000000    0.2701944   0.3205853
## Petal.Length    0.6863769   0.2701944    1.0000000   0.6779013
## Petal.Width     0.3524909   0.3205853    0.6779013   1.0000000
```

```
as_tibble(iris) %>% filter(Species == "setosa") %>% select(-5) %>% cor()
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

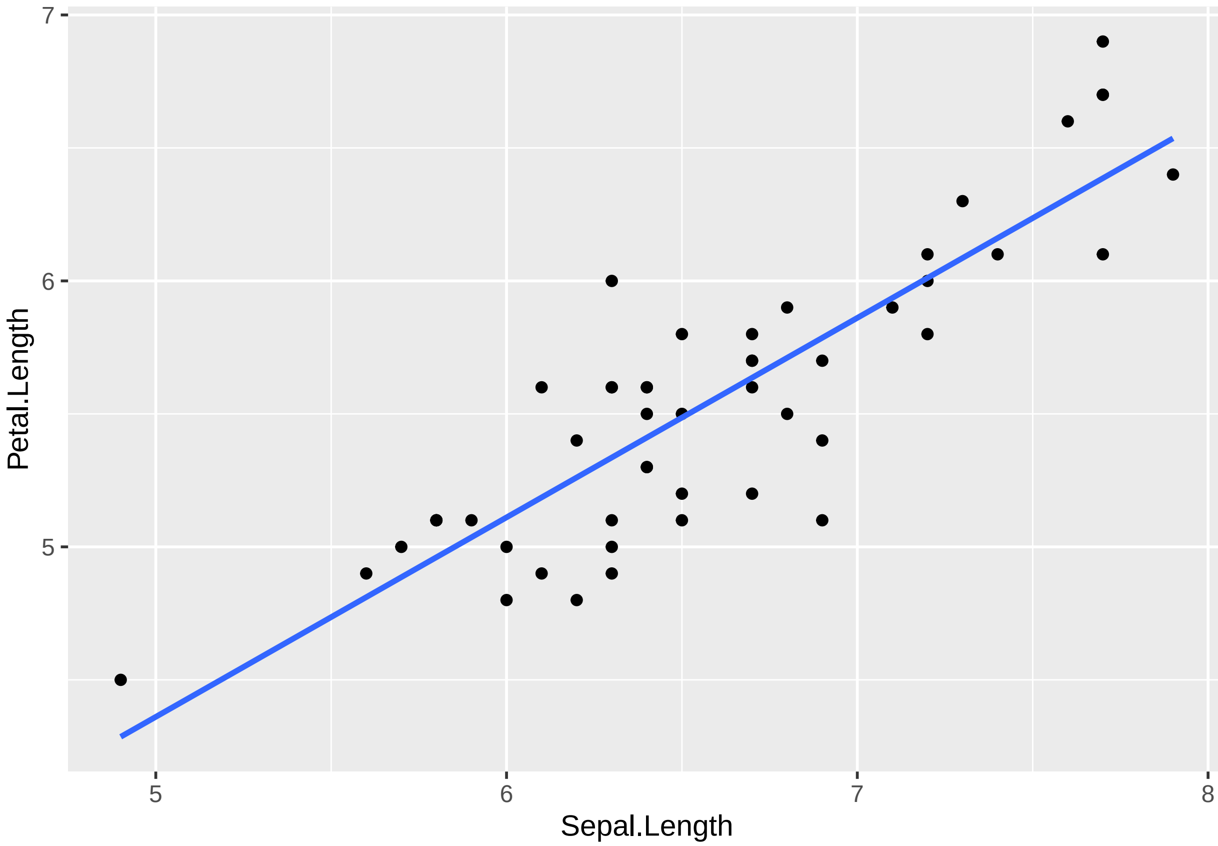
```
as_tibble(iris) %>% filter(Species == "virginica") %>% select(-5) %>% cor()
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.4572278    0.8642247   0.2811077
## Sepal.Width     0.4572278   1.0000000    0.4010446   0.5377280
## Petal.Length    0.8642247   0.4010446    1.0000000   0.3221082
## Petal.Width     0.2811077   0.5377280    0.3221082   1.0000000
```

```
as_tibble(iris) %>% filter(Species == "versicolor") %>% select(-5) %>% cor()
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.5259107    0.7540490   0.5464611
## Sepal.Width     0.5259107   1.0000000    0.5605221   0.6639987
## Petal.Length    0.7540490   0.5605221    1.0000000   0.7866681
## Petal.Width     0.5464611   0.6639987    0.7866681   1.0000000
```

```
as_tibble(iris) %>% filter(Species == "virginica") %>% ggplot(aes(Sepal.Length, Petal.Length)) + geom_p
```



```
as_tibble(iris) %>% filter(Species == "virginica") %>% lm(Petal.Length ~ Sepal.Length, .) %>% glance() %
```

```
## [1] 0.8642247
```

Correlations of the data suggest the possible strength of linear model y ~ x.

```
iris %>% select(-5) %>% cor()
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000  -0.1175698    0.8717538   0.8179411
## Sepal.Width    -0.1175698   1.0000000   -0.4284401  -0.3661259
## Petal.Length    0.8717538  -0.4284401    1.0000000   0.9628654
## Petal.Width     0.8179411  -0.3661259    0.9628654   1.0000000
```

5.1.12 Examples: WDI

- SP.DYN.LE00.IN: Life expectancy at birth, total (years)

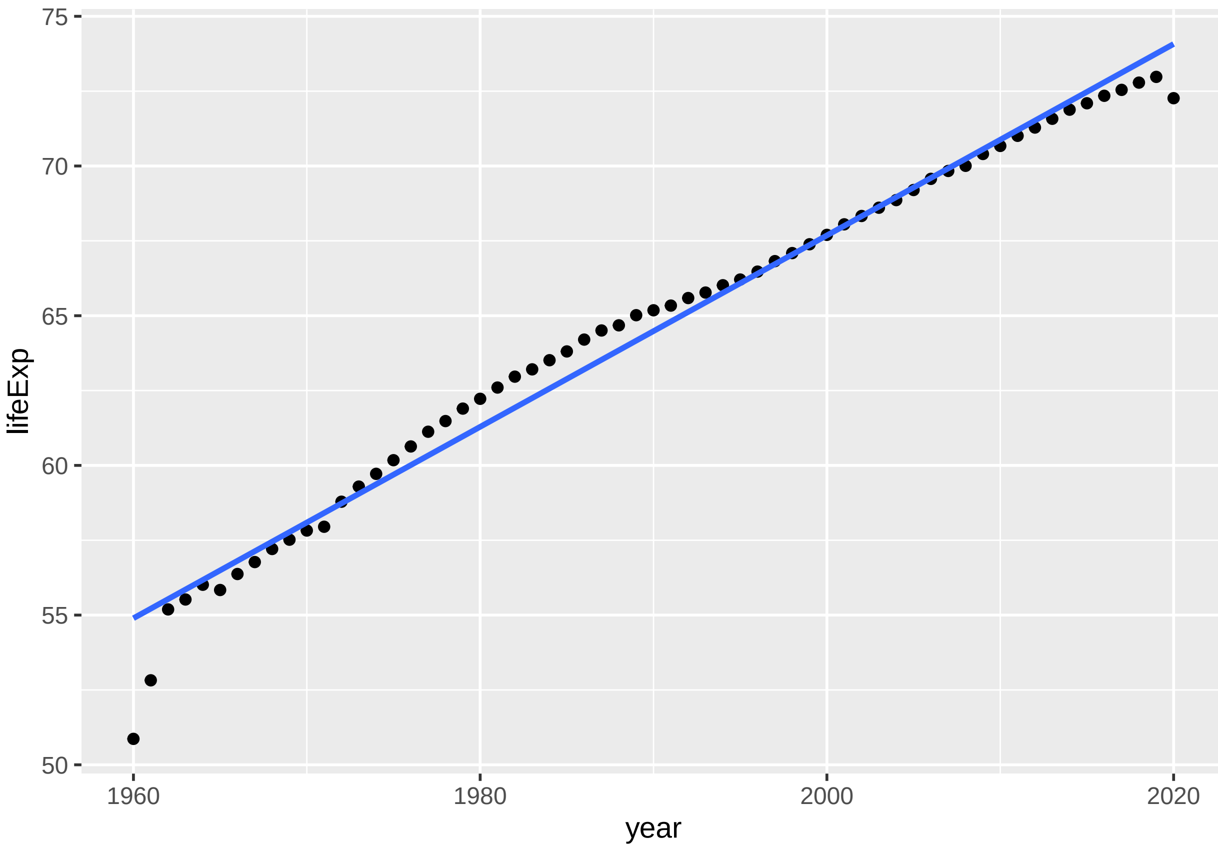
```
wdi_lifeExp <- WDI(indicator = c(lifeExp = "SP.DYN.LE00.IN"))
```

```
## Rows: 16492 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): country, iso2c, iso3c
## dbl (2): year, lifeExp
##
## i Use `spec()` to retrieve the full column specification for this data.
```

```
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
wdi_lifeExp %>% filter(country == "World") %>% drop_na(lifeExp) %>%
  ggplot(aes(year, lifeExp)) + geom_point() + geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
wdi_lifeExp %>% lm(lifeExp ~ year, .) %>% summary()
```

```
##
## Call:
## lm(formula = lifeExp ~ year, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.142  -7.297   1.782   7.873  19.139
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.574e+02  8.987e+00  -62.02   <2e-16 ***
## year         3.123e-01  4.515e-03   69.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.797 on 15202 degrees of freedom
```

```
##   (1288 observations deleted due to missingness)
## Multiple R-squared:  0.2393, Adjusted R-squared:  0.2392
## F-statistic:  4782 on 1 and 15202 DF,  p-value: < 2.2e-16
```

$$lifeExp \sim -557.4 + 0.3123 \cdot year$$

Each year, life expectancy at birth increases approximately 0.3123 years. R-squared of this model is 0.2392, and the model explains 24%.

```r
wdi_lifeExp %>% filter(country == "World", year >= 1962, year <= 2019) %>% drop_na(lifeExp) %>% lm(lifeE
```

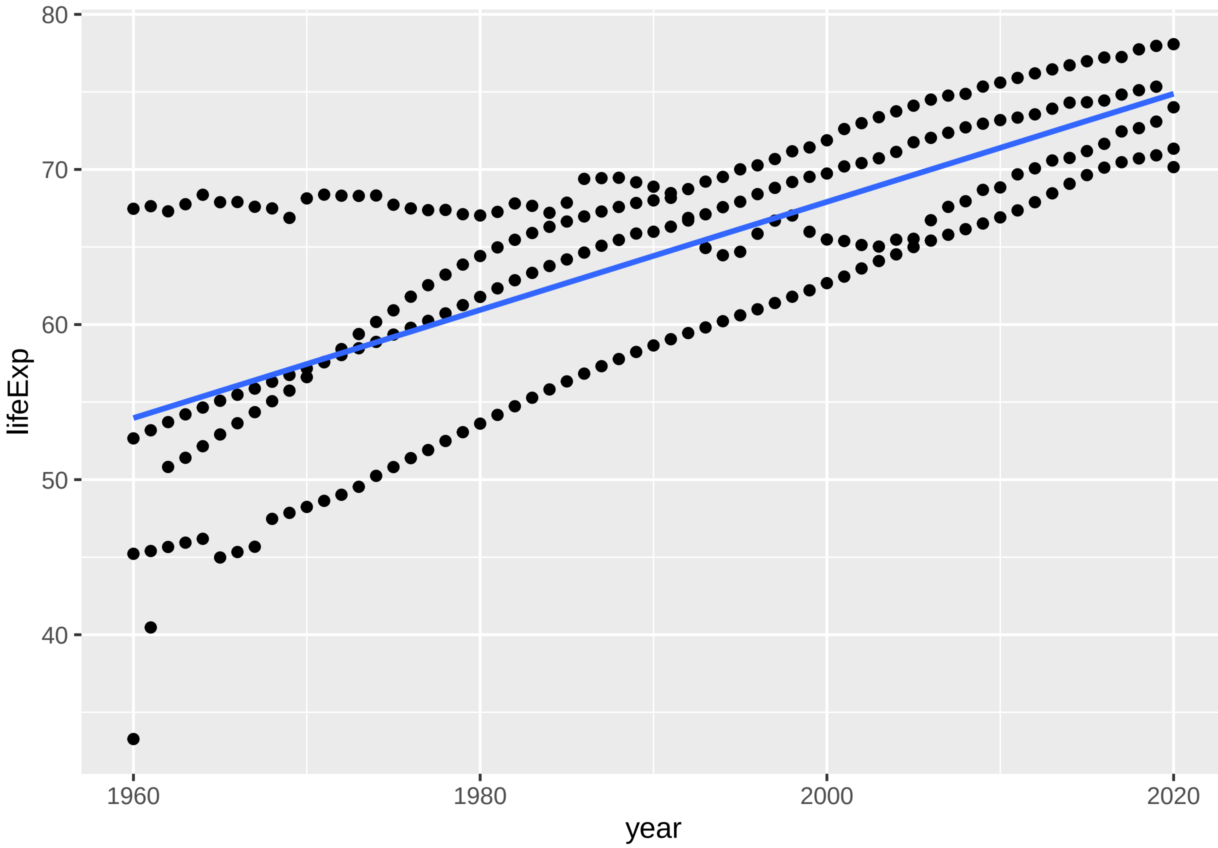
```
##
## Call:
## lm(formula = lifeExp ~ year, data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01769 -0.29535 -0.04302  0.38542  0.82106
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.543e+02  7.885e+00  -70.30   <2e-16 ***
## year         3.110e-01  3.961e-03   78.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.505 on 56 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9908
## F-statistic:  6166 on 1 and 56 DF,  p-value: < 2.2e-16
```

5.1.13 BRICs

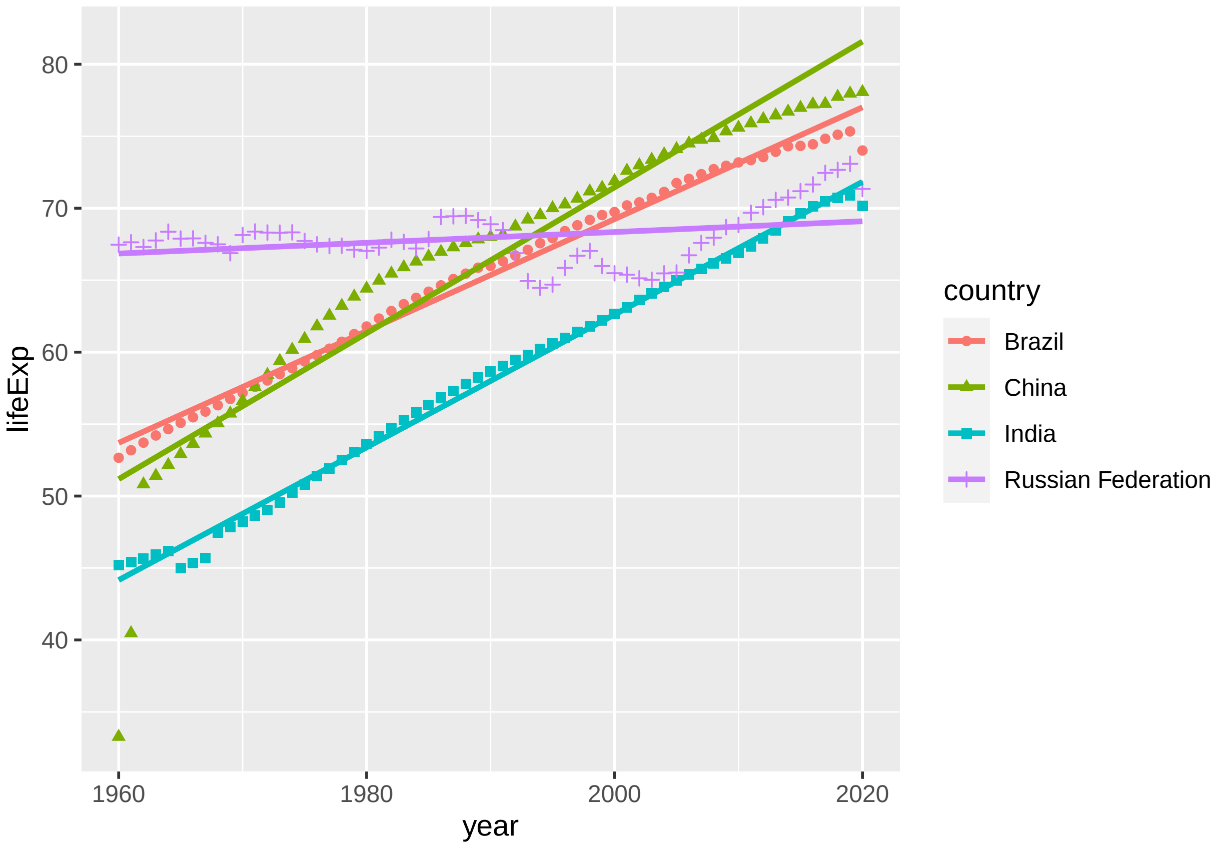
```r
mod_brics <- wdi_lifeExp %>% filter(country %in% c("Brazil", "Russian Federation", "India", "China")) %>
mod_brics$r.squared
```

```
## [1] 0.5658162
```

```r
wdi_lifeExp %>% filter(country %in% c("Brazil", "Russian Federation", "India", "China")) %>% drop_na(li
  ggplot(aes(year, lifeExp)) + geom_point() + geom_smooth(formula = y~x, method = "lm", se = FALSE)
```



```
wdi_lifeExp %>% filter(country %in% c("Brazil", "Russian Federation", "India", "China")) %>% drop_na(li
  ggplot(aes(year, lifeExp, color = country)) + geom_point(aes(shape = country)) + geom_smooth(formula =
```




```
country_model <- function(df) {
  lm(lifeExp ~ year, data = df)
}

by_country <- wdi_lifeExp %>% filter(country %in% c("Brazil", "Russian Federation", "India", "China"))

by_country <- by_country %>%
  mutate(model = map(data, country_model))

by_country %>%
  mutate(tidy = map(model, broom::tidy)) %>%
  unnest(tidy)
```

```
## # A tibble: 8 x 8
## # Groups:   country [4]
##   country            data     model  term     estimate std.e~1 statis~2  p.value
##   <chr>              <list>   <list> <chr>       <dbl>   <dbl>    <dbl>    <dbl>
## 1 Brazil             <tibble> <lm>   (Interc~ -7.08e+2 1.06e+1  -66.6   3.05e-57
## 2 Brazil             <tibble> <lm>   year      3.89e-1 5.34e-3   72.8   1.77e-59
## 3 China              <tibble> <lm>   (Interc~ -9.42e+2 4.79e+1  -19.7   1.32e-27
## 4 China              <tibble> <lm>   year      5.07e-1 2.41e-2   21.1   3.88e-29
## 5 India              <tibble> <lm>   (Interc~ -8.60e+2 8.51e+0 -101.    8.46e-68
## 6 India              <tibble> <lm>   year      4.61e-1 4.28e-3  108.    1.83e-69
## 7 Russian Federation <tibble> <lm>   (Interc~ -6.42e+0 2.69e+1   -0.238 8.12e- 1
## 8 Russian Federation <tibble> <lm>   year      3.74e-2 1.35e-2    2.76  7.59e- 3
## # ... with abbreviated variable names 1: std.error, 2: statistic
```

```r
by_country %>%
  mutate(glance = map(model, broom::glance)) %>%
  unnest(glance)
```

```
## # A tibble: 4 x 15
## # Groups:   country [4]
##   country     data     model r.squ~1 adj.r~2 sigma stati~3  p.value    df logLik
##   <chr>       <list>   <lis>   <dbl>   <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl>
## 1 Brazil      <tibble> <lm>    0.989  0.989  0.734  5.30e3 1.77e-59     1  -66.7
## 2 China       <tibble> <lm>    0.883  0.881  3.31   4.44e2 3.88e-29     1 -159.
## 3 India       <tibble> <lm>    0.995  0.995  0.588  1.16e4 1.83e-69     1  -53.2
## 4 Russian Fe~ <tibble> <lm>    0.115  0.0997 1.86   7.64e0 7.59e- 3     1 -123.
## # ... with 5 more variables: AIC <dbl>, BIC <dbl>, deviance <dbl>,
## #   df.residual <int>, nobs <int>, and abbreviated variable names 1: r.squared,
## #   2: adj.r.squared, 3: statistic
```

5.1.14 Government Expenditure, (% of GDP)

```r
wdi_cache <- read_rds("./data/wdi_cache.RData")
```

```r
WDIsearch("expenditure", "name", cache = wdi_cache) %>%
  inner_join(WDIsearch("% of GDP", "name", cache = wdi_cache))
```

```
## Joining, by = c("indicator", "name")
```

```
##                indicator
## 1     GB.XPD.DEFN.GDP.ZS
## 2      GB.XPD.RSDV.GD.ZS
## 3      GB.XPD.TOTL.GD.ZS
## 4     GB.XPD.TOTL.GDP.ZS
## 5      IE.ICT.TOTL.GD.ZS
## 6      MS.MIL.XPND.GD.ZS
## 7         NE.CON.GOVT.ZS
## 8         NE.CON.PETC.ZS
## 9         NE.CON.PRVT.ZS
## 10        NE.CON.TETC.ZS
## 11        NE.CON.TOTL.ZS
## 12        NE.DAB.TOTL.ZS
## 13     NY.GEN.AEDU.GD.ZS
## 14     SE.XPD.PRIM.PC.ZS
## 15     SE.XPD.SECO.PC.ZS
## 16     SE.XPD.TERT.PC.ZS
## 17     SE.XPD.TOTL.GD.ZS
## 18     SH.XPD.CHEX.GD.ZS
## 19     SH.XPD.GHED.GD.ZS
## 20     SH.XPD.KHEX.GD.ZS
## 21        SH.XPD.PRIV.ZS
## 22        SH.XPD.PUBL.ZS
## 23        SH.XPD.TOTL.ZS
## 24      UIS.XGDP.0.FSGOV
## 25      UIS.XGDP.1.FSGOV
## 26     UIS.XGDP.23.FSGOV
## 27 UIS.XGDP.2T4.V.FSGOV
```

```
## 28     UIS.XGDP.4.FSGOV
## 29    UIS.XGDP.56.FSGOV
##
## 1                                                                                      Defense expenditure (% 
## 2                                                                     Research and development expenditure (% 
## 3                                                                                    Expenditure, total (% 
## 4                                                                                     Total expenditure (% 
## 5                                                   Information and communication technology expenditure (% 
## 6                                                                                   Military expenditure (% 
## 7                                                       General government final consumption expenditure (% 
## 8                                                          Household final consumption expenditure, etc. (% 
## 9                                                    Households and NPISHs final consumption expenditure (% 
## 10                                                                 Final consumption expenditure, etc. (% 
## 11                                                                       Final consumption expenditure (% 
## 12                                                                          Gross national expenditure (% 
## 13                                                            Genuine savings: education expenditure (% 
## 14                                               Government expenditure per student, primary (% of GDP per 
## 15                                             Government expenditure per student, secondary (% of GDP per 
## 16                                              Government expenditure per student, tertiary (% of GDP per 
## 17                                                           Government expenditure on education, total (% 
## 18                                                                          Current health expenditure (% 
## 19                                                    Domestic general government health expenditure (% 
## 20                                                                          Capital health expenditure (% 
## 21                                                                         Health expenditure, private (% 
## 22                                                                          Health expenditure, public (% 
## 23                                                                           Health expenditure, total (% 
## 24                                                  Government expenditure on pre-primary education as % of 
## 25                                                      Government expenditure on primary education as % of 
## 26                                                    Government expenditure on secondary education as % of 
## 27 Government expenditure on secondary and post-secondary non-tertiary vocational education as % of 
## 28                         Government expenditure on post-secondary non-tertiary education as % of 
## 29                                                     Government expenditure on tertiary education as % of 
```

```r
wdi_cache$series %>% filter(grepl("expenditure", name), grepl("% of GDP", name))
```

```
##              indicator
## 1    GB.XPD.DEFN.GDP.ZS
## 2     GB.XPD.RSDV.GD.ZS
## 3    GB.XPD.TOTL.GDP.ZS
## 4     IE.ICT.TOTL.GD.ZS
## 5     MS.MIL.XPND.GD.ZS
## 6        NE.CON.GOVT.ZS
## 7        NE.CON.PETC.ZS
## 8        NE.CON.PRVT.ZS
## 9        NE.CON.TETC.ZS
## 10       NE.CON.TOTL.ZS
## 11       NE.DAB.TOTL.ZS
## 12    NY.GEN.AEDU.GD.ZS
## 13    SE.XPD.PRIM.PC.ZS
## 14    SE.XPD.SECO.PC.ZS
## 15    SE.XPD.TERT.PC.ZS
## 16    SE.XPD.TOTL.GD.ZS
## 17    SH.XPD.CHEX.GD.ZS
```

```
## 18    SH.XPD.GHED.GD.ZS
## 19    SH.XPD.KHEX.GD.ZS
## 20       SH.XPD.PRIV.ZS
## 21       SH.XPD.PUBL.ZS
## 22       SH.XPD.TOTL.ZS
## 23     UIS.XGDP.0.FSGOV
## 24     UIS.XGDP.1.FSGOV
## 25    UIS.XGDP.23.FSGOV
## 26 UIS.XGDP.2T4.V.FSGOV
## 27     UIS.XGDP.4.FSGOV
## 28    UIS.XGDP.56.FSGOV
##
## 1                                                                            Defense expenditure (% o
## 2                                                           Research and development expenditure (% o
## 3                                                                              Total expenditure (% o
## 4                                               Information and communication technology expenditure (% o
## 5                                                                           Military expenditure (% o
## 6                                                   General government final consumption expenditure (% o
## 7                                                      Household final consumption expenditure, etc. (% o
## 8                                              Households and NPISHs final consumption expenditure (% o
## 9                                                              Final consumption expenditure, etc. (% o
## 10                                                                   Final consumption expenditure (% o
## 11                                                                      Gross national expenditure (% o
## 12                                                        Genuine savings: education expenditure (% o
## 13                                            Government expenditure per student, primary (% of GDP per o
## 14                                          Government expenditure per student, secondary (% of GDP per o
## 15                                           Government expenditure per student, tertiary (% of GDP per o
## 16                                                      Government expenditure on education, total (% o
## 17                                                                      Current health expenditure (% o
## 18                                                     Domestic general government health expenditure (% o
## 19                                                                      Capital health expenditure (% o
## 20                                                                     Health expenditure, private (% o
## 21                                                                      Health expenditure, public (% o
## 22                                                                       Health expenditure, total (% o
## 23                                             Government expenditure on pre-primary education as % of G
## 24                                                 Government expenditure on primary education as % of G
## 25                                               Government expenditure on secondary education as % of G
## 26 Government expenditure on secondary and post-secondary non-tertiary vocational education as % of G
## 27                          Government expenditure on post-secondary non-tertiary education as % of G
## 28                                                Government expenditure on tertiary education as % of G
##
## 1
## 2
## 3
## 4
## 5  Military expenditures data from SIPRI are derived from the NATO definition, which includes all cur
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
```

```
## 14
## 15
## 16
## 17
## 18
## 19
## 20
## 21
## 22
## 23
## 24
## 25
## 26
## 27
## 28
##                               sourceDatabase
## 1                       WDI Database Archives
## 2                World Development Indicators
## 3                       WDI Database Archives
## 4               Africa Development Indicators
## 5                World Development Indicators
## 6                World Development Indicators
## 7                       WDI Database Archives
## 8                World Development Indicators
## 9                       WDI Database Archives
## 10               World Development Indicators
## 11               World Development Indicators
## 12                      WDI Database Archives
## 13               World Development Indicators
## 14               World Development Indicators
## 15               World Development Indicators
## 16               World Development Indicators
## 17               World Development Indicators
## 18               World Development Indicators
## 19 Health Nutrition and Population Statistics
## 20                      WDI Database Archives
## 21                      WDI Database Archives
## 22                      WDI Database Archives
## 23                       Education Statistics
## 24                       Education Statistics
## 25                       Education Statistics
## 26                       Education Statistics
## 27                       Education Statistics
## 28                       Education Statistics
##
## 1
## 2  UNESCO Institute for Statistics (UIS). UIS.Stat Bulk Data Download Service. Accessed October 24, 
## 3
## 4                    World Information Technology and Services Alliance, Digital Planet: The Global I
## 5                          Stockholm International Peace Research Institute (SIPRI), Yearbook: Armame
## 6                                                                              World Bank national accounts 
## 7                                                                              World Bank national accounts 
## 8                                                                              World Bank national accounts 
## 9                                                                              World Bank national accounts 
```

```
## 10                                                                   World Bank national accounts d
## 11                                                                   World Bank national accounts d
## 12
## 13                                                        UNESCO Institute for Statistics (http:/
## 14                                                        UNESCO Institute for Statistics (http:/
## 15                                                        UNESCO Institute for Statistics (http:/
## 16 UNESCO Institute for Statistics (UIS). UIS.Stat Bulk Data Download Service. Accessed October 24, 2
## 17  World Health Organization Global Health Expenditure database (http://apps.who.int/nha/database).
## 18  World Health Organization Global Health Expenditure database (http://apps.who.int/nha/database).
## 19  World Health Organization Global Health Expenditure database (http://apps.who.int/nha/database).
## 20             World Health Organization Global Health Expenditure database (see http://apps.who.int
## 21             World Health Organization Global Health Expenditure database (see http://apps.who.int
## 22             World Health Organization Global Health Expenditure database (see http://apps.who.int
## 23
## 24
## 25
## 26
## 27
## 28
```

- NY.GDP.PCAP.KD: GDP per capita (constant 2015 US$)
- SP.DYN.LE00.IN: Life expectancy at birth, total (years)
- SP.POP.TOTL: Population, total
- GB.XPD.RSDV.GD.ZS: Research and development expenditure (% of GDP) - 2
- MS.MIL.XPND.GD.ZS: Military expenditure (% of GDP) - 6
- SE.XPD.TOTL.GD.ZS: Government expenditure on education, total (% of GDP)

```r
wdi_world <- WDI(country = "all", indicator = c(gdpPcap = "NY.GDP.PCAP.KD", lifeExp = "SP.DYN.LE00.IN",
```

```
## Rows: 8512 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (7): country, iso2c, iso3c, region, capital, income, lending
## dbl  (9): year, gdpPcap, lifeExp, pop, research, military, education, longit...
## lgl  (1): status
## date (1): lastupdated
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
wdi_world
```

```
## # A tibble: 8,512 x 18
##    country    iso2c iso3c  year status lastupda~1 gdpPcap lifeExp    pop resea~2
##    <chr>      <chr> <chr> <dbl> <lgl>  <date>       <dbl>   <dbl>  <dbl>   <dbl>
##  1 Afghanist~ AF    AFG    2018 NA     2022-12-22    579.    63.1 3.67e7      NA
##  2 Afghanist~ AF    AFG    2009 NA     2022-12-22    512.    60.4 2.74e7      NA
##  3 Afghanist~ AF    AFG    2016 NA     2022-12-22    590.    63.1 3.46e7      NA
##  4 Afghanist~ AF    AFG    2014 NA     2022-12-22    603.    62.5 3.27e7      NA
##  5 Afghanist~ AF    AFG    2012 NA     2022-12-22    596.    61.9 3.05e7      NA
##  6 Afghanist~ AF    AFG    2015 NA     2022-12-22    592.    62.7 3.38e7      NA
##  7 Afghanist~ AF    AFG    1990 NA     2022-12-22     NA     46.0 1.07e7      NA
```

```
##  8 Afghanist~ AF    AFG    2019 NA     2022-12-22    584.    63.6 3.78e7      NA
##  9 Afghanist~ AF    AFG    2002 NA     2022-12-22    360.    56.5 2.10e7      NA
## 10 Afghanist~ AF    AFG    2017 NA     2022-12-22    589.    63.0 3.56e7      NA
## # ... with 8,502 more rows, 8 more variables: military <dbl>, education <dbl>,
## #   region <chr>, capital <chr>, longitude <dbl>, latitude <dbl>, income <chr>,
## #   lending <chr>, and abbreviated variable names 1: lastupdated, 2: research
```

SE.XPD.TOTL.GB.ZS: Government expenditure on education, total (% of government expenditure) SE.XPD.TOTL.GD.ZS: Government expenditure on education, total (% of GDP) SE.XPD.PRIM.PC.ZS: Government expenditure per student, primary (% of GDP per capita) MS.MIL.XPND.ZS: Military expenditure (% of general government expenditure) SE.XPD.TERT.ZS: Expenditure on tertiary education (% of government expenditure on education) —

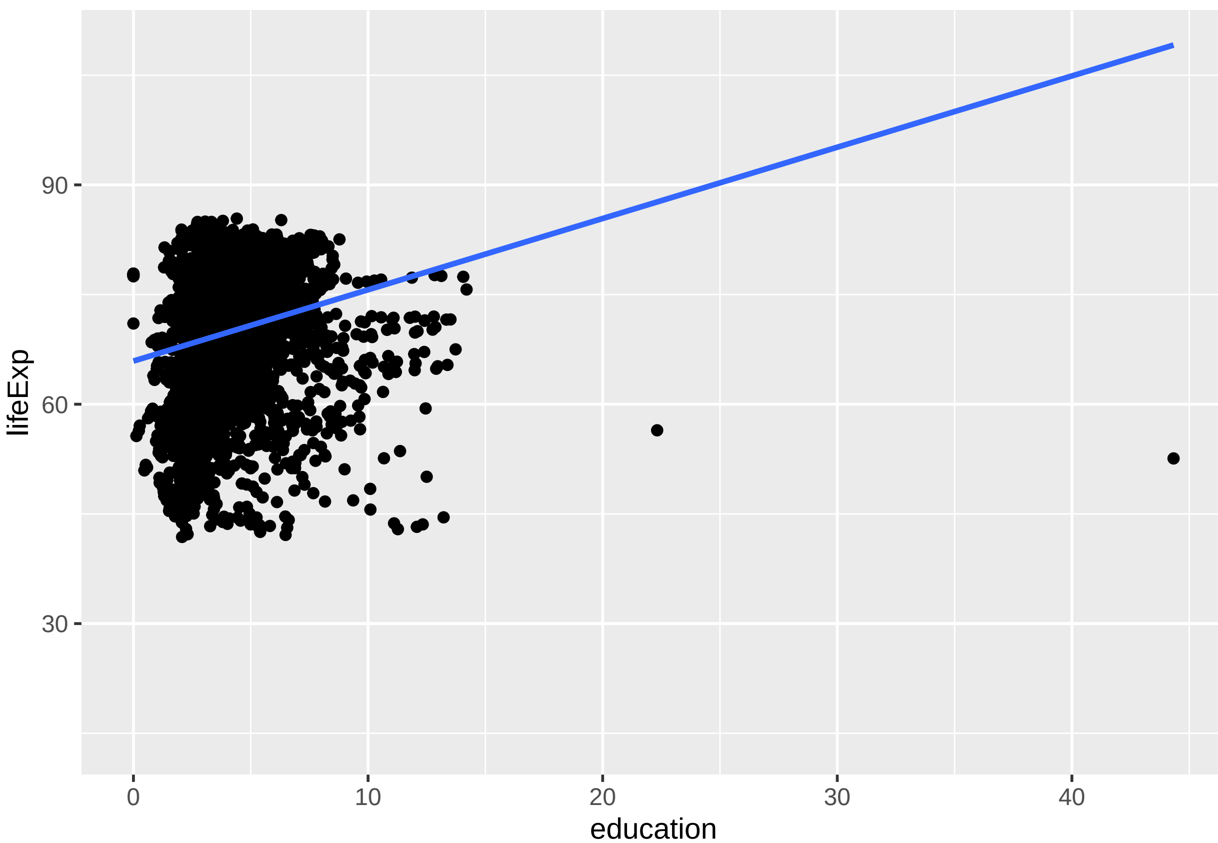
```
mod_e <- lm(lifeExp ~ education, wdi_world); mod_e
```

```
##
## Call:
## lm(formula = lifeExp ~ education, data = wdi_world)
##
## Coefficients:
## (Intercept)    education
##     65.9047       0.9748
```

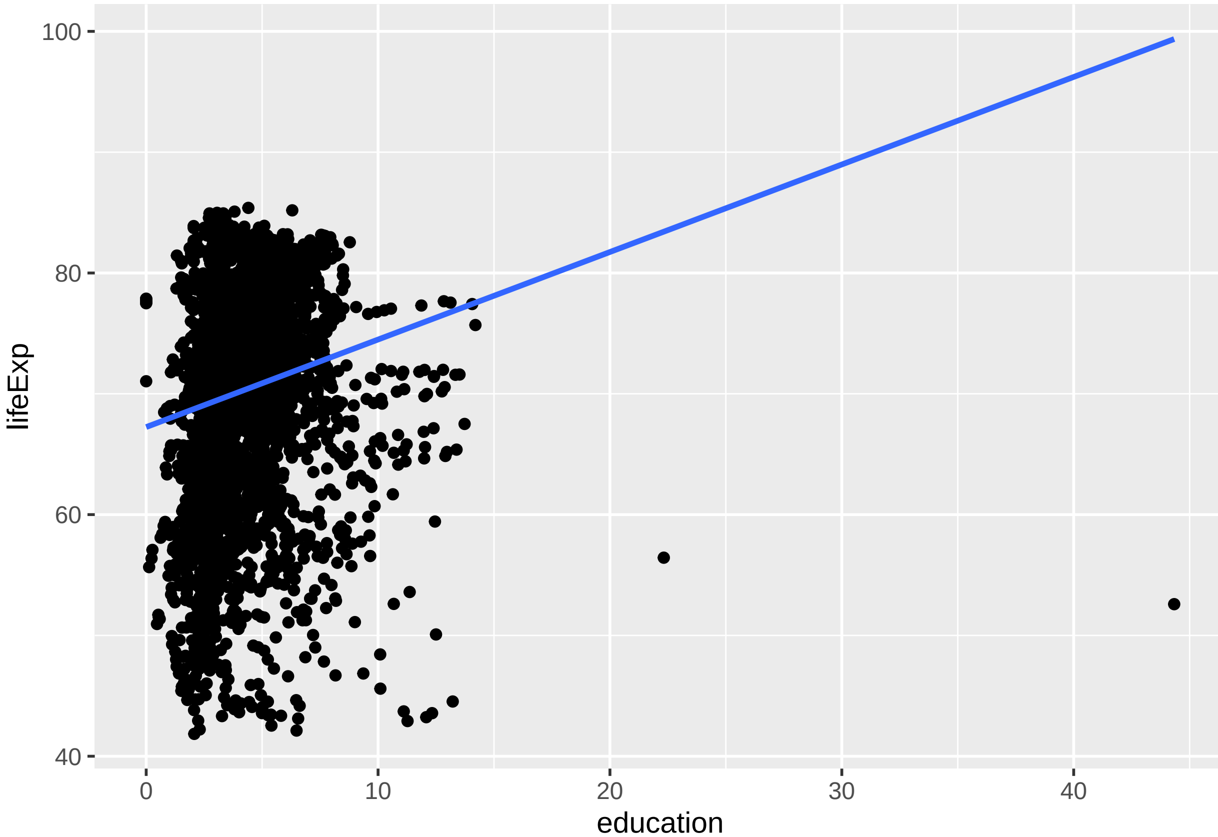
```
wdi_world %>% ggplot(aes(education, lifeExp)) + geom_point() + geom_smooth(formula = y ~ x, method = "lm
```

```
## Warning: Removed 3858 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 3858 rows containing missing values (`geom_point()`).
```



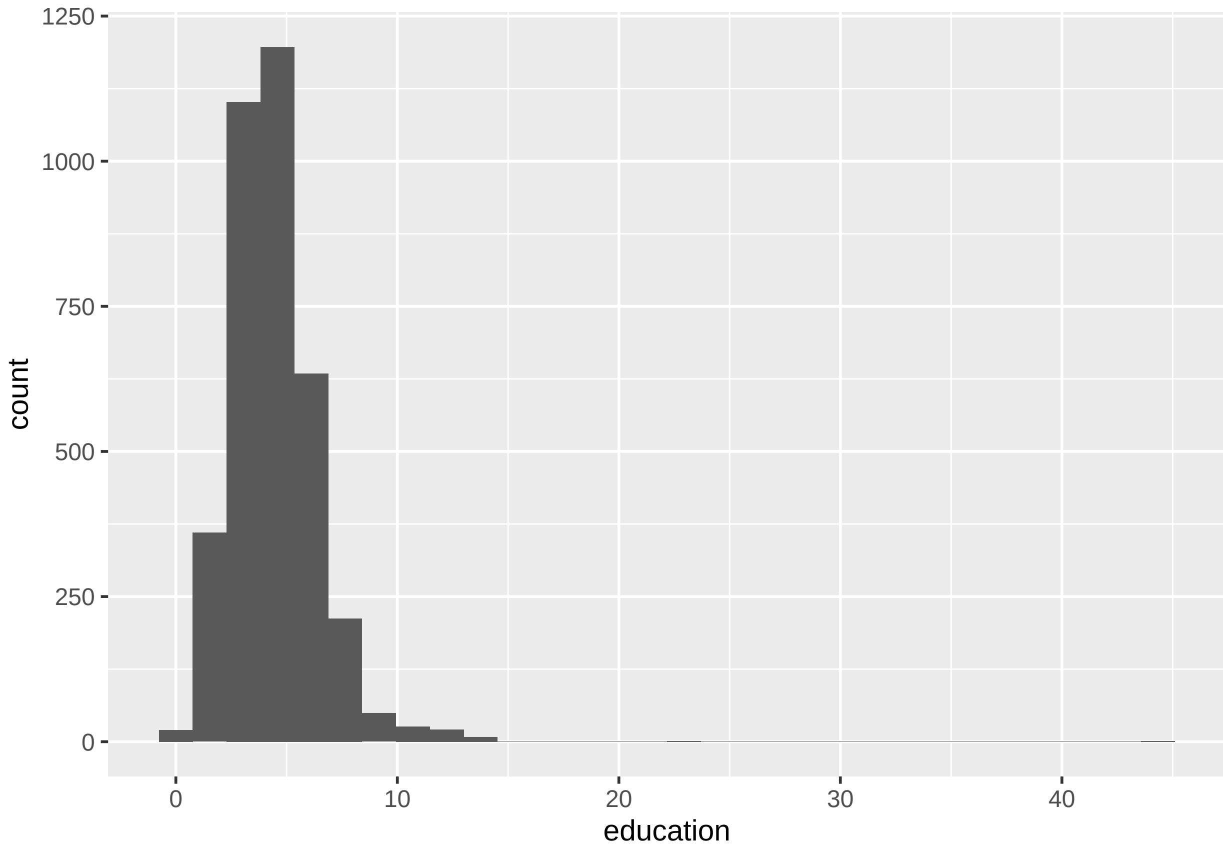
```
wdi_world %>% filter(income != "Aggregates") %>% drop_na(education, lifeExp) %>% ggplot(aes(education, l
```



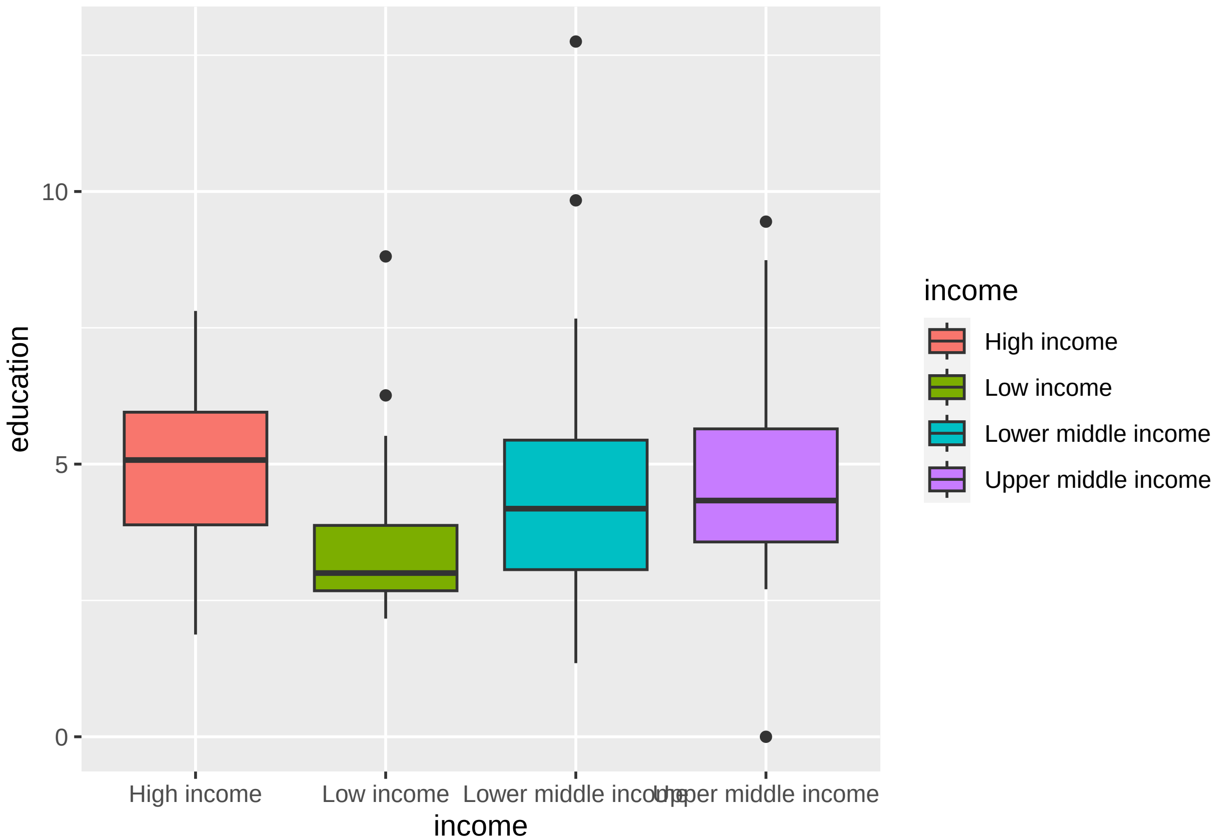

```
wdi_world_el <- wdi_world %>% select(country, year, education, lifeExp, gdpPcap, pop, research, militar
```

```
wdi_world_el %>% ggplot(aes(education)) + geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
wdi_world_el %>% filter(year==2020) %>% ggplot(aes(x = income, y = education, fill = income)) + geom_bo
```




```
wdi_world_el %>% filter(year==2020) %>% arrange(desc(education))
```

```
## # A tibble: 152 x 10
##    country      year educa~1 lifeExp gdpPcap    pop resea~2 milit~3 region income
##    <chr>       <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <chr>  <chr>
##  1 Solomon I~   2020   12.8     70.2   2080. 6.91e5  NA      NA     East ~ Lower~
##  2 Bolivia      2020    9.84    64.5   2920. 1.19e7  NA       1.32  Latin~ Lower~
##  3 Namibia      2020    9.45    62.8   4155. 2.49e6  NA       3.23  Sub-S~ Upper~
##  4 Sierra Le~   2020    8.81    59.8    604. 8.23e6  NA       0.547 Sub-S~ Low i~
##  5 Botswana     2020    8.74    65.6   5811. 2.55e6  NA       3.20  Sub-S~ Upper~
##  6 Saudi Ara~   2020    7.81    76.2  18086. 3.60e7   0.522   9.22  Middl~ High ~
##  7 Iceland      2020    7.72    83.1  52984. 3.66e5   2.47   NA     Europ~ High ~
##  8 Lesotho      2020    7.67    54.7    972. 2.25e6  NA       1.62  Sub-S~ Lower~
##  9 Cabo Verde   2020    7.58    74.8   2801. 5.83e5  NA       0.590 Sub-S~ Lower~
## 10 Belize       2020    7.53    72.9   5040. 3.95e5  NA       1.72  Latin~ Upper~
## # ... with 142 more rows, and abbreviated variable names 1: education,
## #   2: research, 3: military
```

```
wdi_world_el %>% filter(year==2020) %>% arrange(desc(education))
```

```
## # A tibble: 152 x 10
##    country      year educa~1 lifeExp gdpPcap    pop resea~2 milit~3 region income
##    <chr>       <dbl>   <dbl>   <dbl>   <dbl>  <dbl>   <dbl>   <dbl> <chr>  <chr>
##  1 Solomon I~   2020   12.8     70.2   2080. 6.91e5  NA      NA     East ~ Lower~
##  2 Bolivia      2020    9.84    64.5   2920. 1.19e7  NA       1.32  Latin~ Lower~
##  3 Namibia      2020    9.45    62.8   4155. 2.49e6  NA       3.23  Sub-S~ Upper~
##  4 Sierra Le~   2020    8.81    59.8    604. 8.23e6  NA       0.547 Sub-S~ Low i~
##  5 Botswana     2020    8.74    65.6   5811. 2.55e6  NA       3.20  Sub-S~ Upper~
```

```
##  6 Saudi Ara~  2020    7.81    76.2  18086. 3.60e7   0.522   9.22  Middl~ High ~
##  7 Iceland     2020    7.72    83.1  52984. 3.66e5   2.47   NA     Europ~ High ~
##  8 Lesotho     2020    7.67    54.7    972. 2.25e6  NA       1.62  Sub-S~ Lower~
##  9 Cabo Verde  2020    7.58    74.8   2801. 5.83e5  NA       0.590 Sub-S~ Lower~
## 10 Belize      2020    7.53    72.9   5040. 3.95e5  NA       1.72  Latin~ Upper~
## # ... with 142 more rows, and abbreviated variable names 1: education,
## #   2: research, 3: military
```

```
wdi_world_el %>% filter(year==2020) %>% lm(gdpPcap ~ education, .)
```

```
##
## Call:
## lm(formula = gdpPcap ~ education, data = .)
##
## Coefficients:
## (Intercept)    education
##        9158         1285
```

```
wdi_world_el %>% filter(year==2020) %>% lm(gdpPcap ~ education, .) %>% glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1  sigma stati~2 p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl>  <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1    0.0131      0.00650 20523.    1.98   0.161     1 -1713. 3431. 3440. 6.28e10
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
wdi_world_el %>% lm(lifeExp ~ education + research + military, .) %>% glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squ~1 sigma stati~2   p.value    df logLik   AIC   BIC devia~3
##       <dbl>       <dbl> <dbl>   <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.346       0.345  5.25    270. 2.05e-140     3 -4711. 9432. 9458.  42036.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

```
wdi_world_el %>% lm(lifeExp ~ education + research + military, .) %>% tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  70.2       0.489    144.    0
## 2 education     0.0771    0.0966     0.798 4.25e-  1
## 3 research      3.84      0.145     26.4   6.95e-127
## 4 military     -0.0682    0.102     -0.667 5.05e-  1
```

$$lifeExp \sim 70.22 + 0.08 \cdot education + 3.84 \cdot research - 0.07 \cdot military$$

```
wdi_world_el %>% lm(gdpPcap ~ education + research + military, .) %>% tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    1077.     1308.     0.823 4.11e-  1
## 2 education      1324.      258.     5.12  3.41e-  7
## 3 research      12792.      389.    32.9   1.07e-179
## 4 military       -967.      273.    -3.54  4.08e-  4
```

```
wdi_world_el %>% lm(gdpPcap ~ education + research + military, .) %>% glance()
```

```
## # A tibble: 1 x 12
##   r.squared adj.r~1  sigma stati~2   p.value    df  logLik    AIC    BIC devia~3
##       <dbl>   <dbl>  <dbl>   <dbl>     <dbl> <dbl>   <dbl>  <dbl>  <dbl>   <dbl>
## 1     0.478   0.477 14013.    466. 9.65e-215     3 -16766. 33542. 33569. 2.99e11
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

$$gdpPcap \sim 1077 + 1024 \cdot education + 12792 \cdot research - 967 \cdot military$$

```
mod_r <- lm(lifeExp ~ research, wdi_world); mod_e
```

```
##
## Call:
## lm(formula = lifeExp ~ education, data = wdi_world)
##
## Coefficients:
## (Intercept)    education
##     65.9047       0.9748
```

5.1.15 model and Linear Regression Quick Reference

- R4DS: Model basics
  - https://r4ds.had.co.nz/model-basics.html

For explanation of other indices, please see.

- r-statistics.co by Selva Prabhakaran:
  - http://r-statistics.co/Linear-Regression.html

5.2 Roudups

5.2.1 R Markdown Revisited

Presentation: Submit an R Notebook (with codes used in the presentation), and PowerPoint file or other files used for your presentation, if any. If you use R Notebook for your presentation, you do not need to submit extra files.

Final Paper: Submit an R Notebook (with codes as a work file), and a PDF (rendered directly from an R Notebook, or created from Word) - Maximum pages of PDF is eight.

Format of Presentation - R Notebook is fine and slide presentation in various format is also fine

---

5.2.1.1 Literate Programming and Reproducible Research Importing Data:

1. Read a csv file: `read_csv("./data/file_name.csv")`
2. Download and import using a url of a csv file: `read_csv(url)`
3. Read an Excel file: `readxl::read_excel("./data/excel_file_name.xlsx")`
4. Read from the clipboard: `read_delim(clipboard())`

- zip file:

  - copy the url
  - wir1to10 <- "https://wir2022.wid.world/www-site/uploads/2022/03/WIR2022TablesFigures-Chapter.zip"
  - download.file(wir1to10, destfile = "./data/wir1to10.zip")
  - unzip("./data/wir1to10.zip", exdir = "./data")
  - list.files("./data/WIR2022TablesFigures-Chapter")
  - excel_sheets("./data/WIR2022TablesFigures-Chapter/WIR2022TablesFigures-Chapter1.xlsx")
  - df <- read_delim(clipboard()); df
  - Not reproducible unless clearly explained.

5.2.1.2 Code Chunk Options https://yihui.org/knitr/options/

- Chunk Name
- Output: use document default
  - Show code and output: echo=TRUE, eval=TRUE - Default
  - Show output only: echo=FALSE
  - Show nothing (run code): include=FALSE
  - Show nothing (don't run code): include=FALSE, eval=FALSE
- Show message: message=TRUE, FALSE
- Show warning: warning=TRUE, FALSE
- Use Paged Tables: paged.print=TRUE, FALSE
- Use custom figure size: width and height in inch.
- You can use Hide Code and Show Code option on the rendered Notebook file.

5.2.1.3 Presentation and Paper

1. Data Source
2. Variables
3. Problems
4. Visualization
5. Model
6. Conclusions and Further Research

   WDI, WIR, etc

5.2.1.4 Word Custom Word templates: https://bookdown.org/yihui/rmarkdown-cookbook/word-template.html

You can apply the styles defined in a Word template document to new Word documents generated from R Markdown. Such a template document is also called a "style reference document." The key is that you have to create this template document from Pandoc first, and change the style definitions in it later. Then pass the path of this template to the reference_docx option of word_document

```
---
 word_document:
    reference_docx: "template.docx"
---
```

5.2.1.5 PowerPoint

PowerPoint presentation: https://bookdown.org/yihui/rmarkdown/powerpoint-presentation.html

Custom templates: https://bookdown.org/yihui/rmarkdown/powerpoint-presentation.html#ppt-templates

```
---
  powerpoint_presentation:
    reference_doc: my-styles.pptx
---
```

https://support.microsoft.com/en-us/office/create-and-save-a-powerpoint-template-ee4429ad-2a74-4100-82f7-50f8169c8aca

YouTube: How To Create A PowerPoint Template

5.3 The Week Six Assignment - Assignment Five (in Moodle)

- Choose a public data. Clearly state how you obtained the data. Even if you are able to give the URL to download the data, explain the steps you reached and obtained the data.
- Create an R Notebook of a Data Analysis containing the following and submit the rendered HTML file (eg. `a5_123456.nb.html` by replacing 123456 with your ID), and a PDF (or MS Word File).
  1. create an R Notebook using the R Notebook Template in Moodle, save as `a3_123456.Rmd`,
  2. write your name and ID and the contents,
  3. run each code block,
  4. preview to create `a5_123456.nb.html`,
  5. render (or knit) PDF, or Word (and then PDF)
  6. submit `a5_123456.nb.html` and PDF (or Word) to Moodle.

1. Choose a data with at least two numerical variables. One of them can be the year.
   - Information of the data
   - Explain why you chose the data
   - List questions you want to study

---

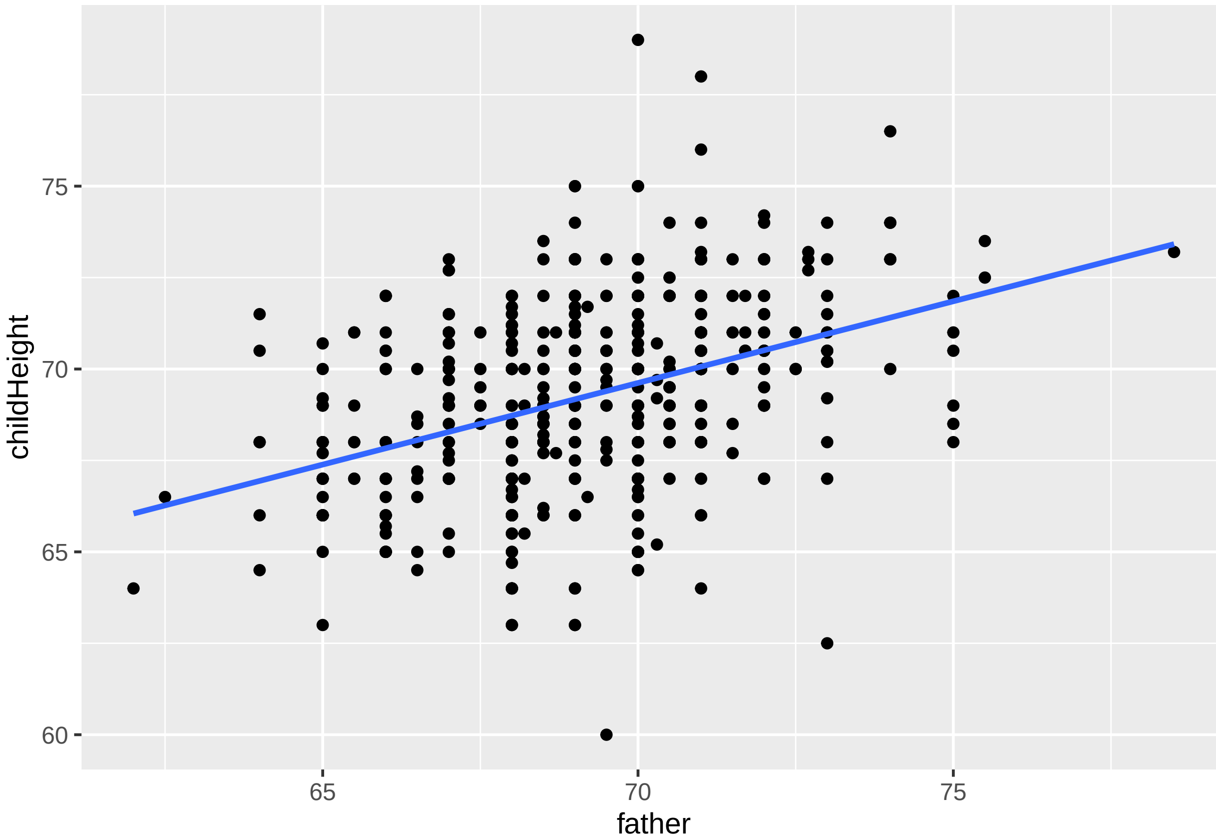
2. Explore the data using visualization using `ggplot2`
   - Create various charts, and write observed comments
   - Apply a (linear regression) model, and draw a regression line to at least one chart, and write your conclusion based on the model using the slope value and R squared (and/or adjusted R squared).
3. Observations based on your data visualization, and difficulties and questions encountered if any.

Due: 2023-01-30 23:59:00. Submit your R Notebook file, and a PDF file (or a MS Word file) in Moodle (The Fifth Assignment). Due on Monday!

5.4 Roundup

5.4.1 History of Regression Analysis: slope = 0.4465

The heights of descendants of tall ancestors tend to regress down towards a normal average



5.4.2 Anna Karenina Principle

> "Tidy data sets are all alike; but every messy data set is messy in its own way." — Hadley Wickham

> "all happy families are all alike; each unhappy family is unhappy in its own way" - Tolstoy's Anna Karenina

The Anna Karenina principle states that a deficiency in any one of a number of factors dooms an endeavor to failure. Consequently, a successful endeavor (subject to this principle) is one for which every possible deficiency has been avoided. (Wikipedia)

Please look at the outliers carefully.