Week 5: Linear regression

Monica Alexander

06/02/2022

By the end of this lab you should know

- How to use `lm` to get estimated coefficients and R^2
- How to calculate an estimated outcome (Y) based on a particular value of an independent variable (X) and estimated regression coefficients
- How to plot a fitted SLR line on a scatter plot

Read in, prepare, plot the data

We will be using the country indicators dataset again, and exploring the relationship between the total fertility rate (TFR) and child mortality in 2017.

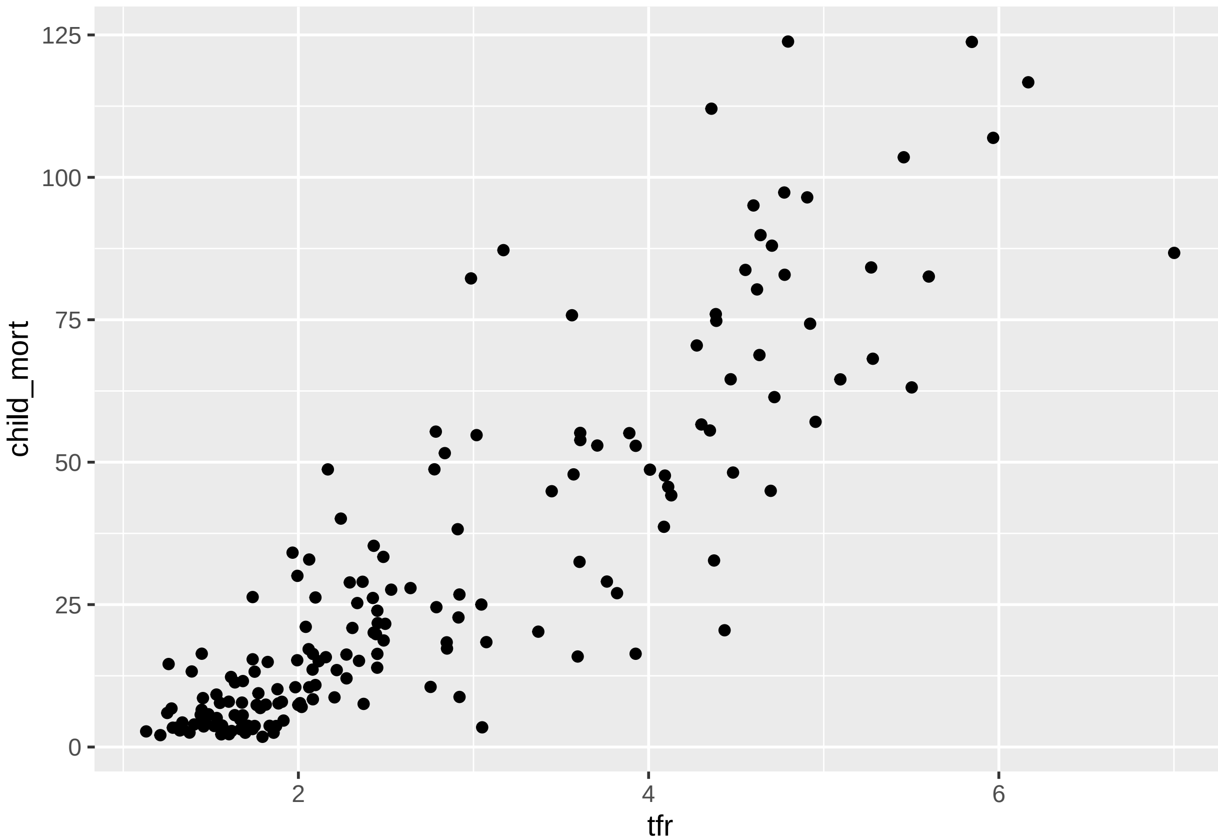
```
library(tidyverse)
library(here)
country_ind <- read_csv(here("data/country_indicators.csv"))

# NOTE if you are having trouble with the 'here' package
# don't use it and just type in the whole file path.
```

Filter to just be 2017

```
country_ind_2017 <- country_ind %>% filter(year==2017)
```

Look at the observed relationship between TFR and child mortality

```
ggplot(country_ind_2017, aes(tfr, child_mort)) +
  geom_point()
```



Question

Alter the code above to make a plot title and make the X and Y axes more readable.

Estimating SLR using `lm`

We don't have to calculate the regression coefficients or R^2 'by hand', we can just use the `lm` function. ('lm' stands for 'linear models').

The main arguments are

- The formula, which is written in the form `y~x`
- The data frame that contains the variables

Fit our SLR:

```
childmort_tfr_model <- lm(formula = child_mort~tfr, data = country_ind_2017)
```

Print out the summary:

```
summary(childmort_tfr_model)
```

```
##
## Call:
## lm(formula = child_mort ~ tfr, data = country_ind_2017)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.937  -7.093  -0.558   5.404  52.029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -25.8338     2.6135  -9.885   <2e-16 ***
## tfr          20.3581     0.8579  23.730   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.73 on 174 degrees of freedom
## Multiple R-squared:  0.7639, Adjusted R-squared:  0.7626
## F-statistic: 563.1 on 1 and 174 DF,  p-value: < 2.2e-16
```

Confirm that the resulting values are the same as the ones you obtained by doing the calculations 'by hand'.

Extract results

To extract coefficients from the model output, use the `coef()` function

```
coef(childmort_tfr_model)
```

```
## (Intercept)         tfr
##   -25.83383    20.35806
```

Can assign these to variables by indexing the relevant number:

```
beta_0 <- coef(childmort_tfr_model)[1] # the [1] means get the first item
beta_1 <- coef(childmort_tfr_model)[2]
```

To extract the value of R^2, use

```
summary(childmort_tfr_model)[["r.squared"]]
```

```
## [1] 0.7639494
```

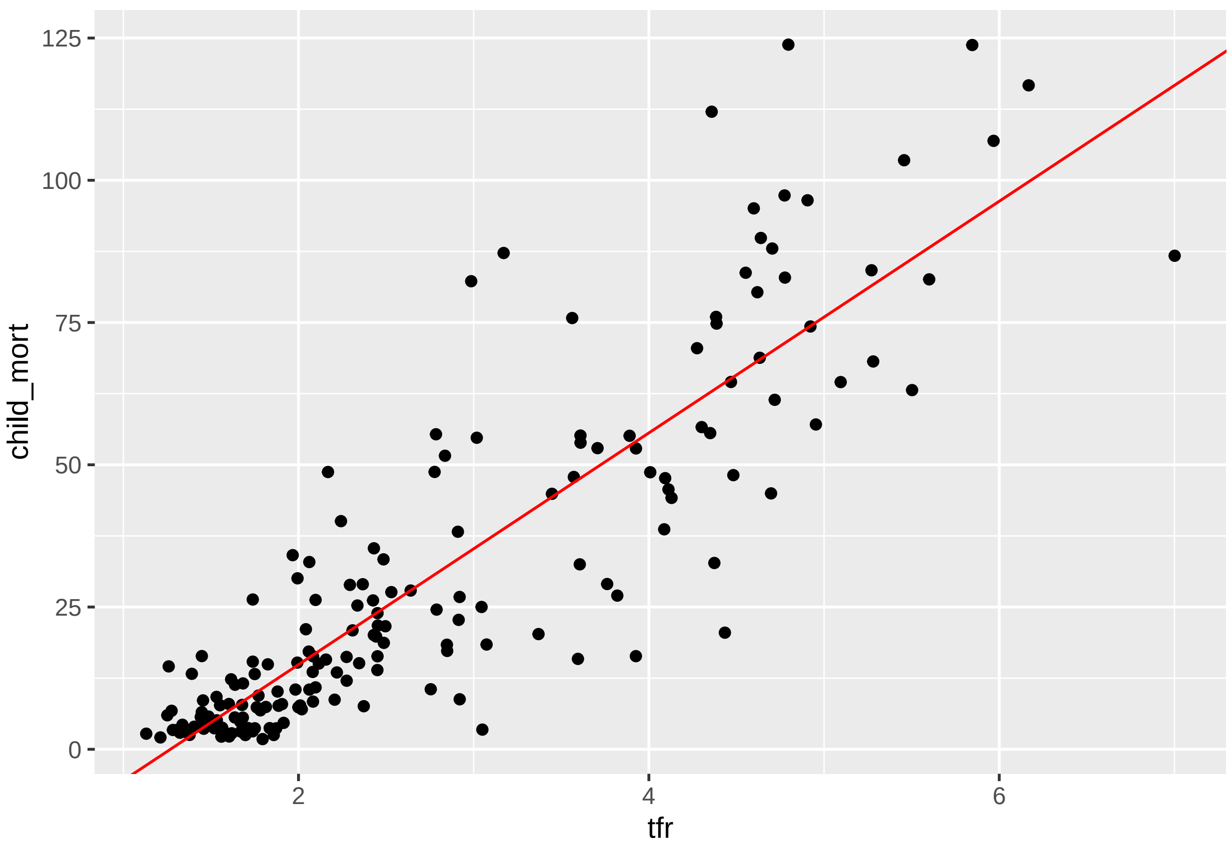
Question

What is the estimated child mortality for a country with a TFR or 5?

Plotting the fitted line on a scatter plot

To visualize our fitted line we can add to our plot from before using `geom_abline`

```r
ggplot(country_ind_2017, aes(tfr, child_mort)) +
  geom_point() +
  geom_abline(intercept = beta_0, slope = beta_1, color = "red")
```



Multiple linear regression

Running MLR in R is an easy extension of SLR. Here are some practice questions using the `lego_towers` dataset. This dataset shows observations of the time (in seconds) it took my toddler to build a lego tower, the number of blocks given to him, and the number of other distractions present.

Reading in the data:

```r
lego <- read_csv(here("data/lego_towers.csv"))
```

Questions

1. Make a scatter plot of time versus blocks
2. Make a scatter plot of time versus distractions
3. Based on 1 and 2, what do you expect the magnitude and sign (direction) of $\hat{\beta}_1$ and $\hat{\beta}_2$ to be?
4. Fit the above model using `lm`
5. Interpret $\hat{\beta}_1$ and $\hat{\beta}_2$
6. Using `mutate` create a new variable called `blocks_3` which is the number of blocks minus 3
7. Refit the model using `lm` where X_{i1} is now `blocks_3`
8. Interpret $\hat{\beta}_0$