Lab 3 Key

Exercises

You may want to create a new quarto file for the exercises. This way, it will be easy to keep track of what you are doing and avoid confusion with the work done during class. For this lab you need the same packages used in class (i.e., openintro, tidyverse, statsr, and broom).

1. ***(2 pts)*** Create a new data frame (name it `new_hfi`) with the variables `region`, `ef_regulation`, `ef_regulation_business` and `ef_score`. You will need this data frame in the next exercises.

```
data(hfi)
```

```
new_hfi<- hfi%>%
  select(region, ef_score, ef_regulation, ef_regulation_business)
```

2. ***(2 pts)*** Create a new data frame (name it `new_hfi_1`) from `new_hfi`. The data frame `new_hfi_1` should have countries from `South Asia` or `East Asia` only. How many cases are in this new data frame?

```
new_hfi_1<- new_hfi%>%
  filter(region== "South Asia"|region== "East Asia")
```

3. ***(2 pts)*** Use the `group_by` function to compute the average `ef_score` for counties in East Asia and those in South Asia.

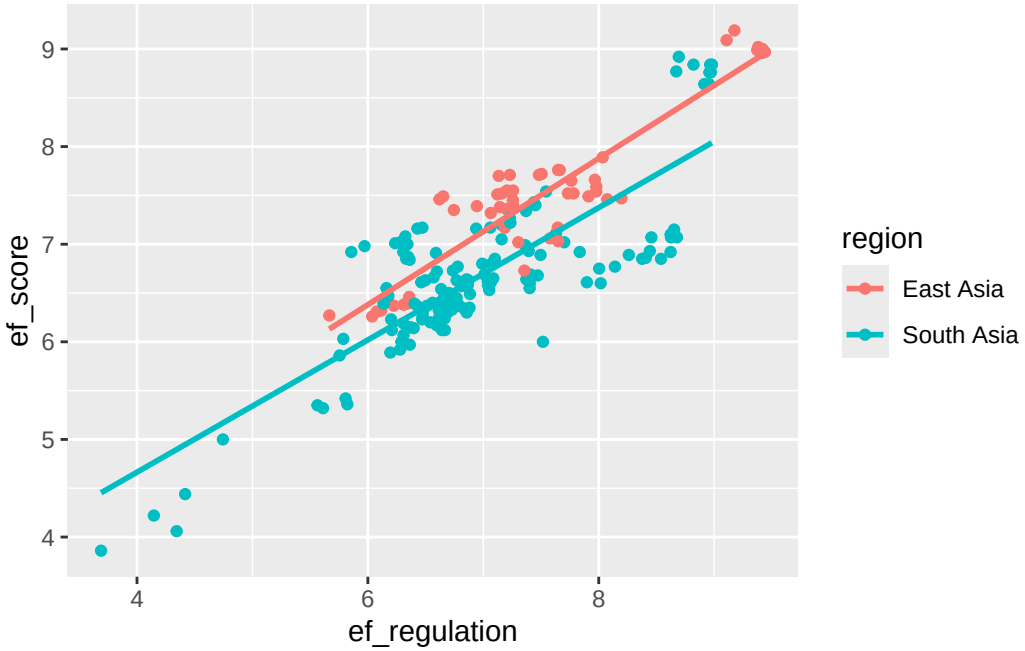
```
new_hfi_1%>%
  drop_na(ef_score)%>%
  group_by(region)%>%
  summarize(mean(ef_score))
```

```
# A tibble: 2 x 2
  region     `mean(ef_score)`
  <chr>                 <dbl>
1 East Asia              7.54
2 South Asia             6.66
```

4. ***(4 pts)*** Recreate the following scatter plot using the `new_hfi_1` data frame. What can you say about the relationship between `ef_score` and `ef_regulation` for countries in East Asia and South Asia?

```
ggplot(data=new_hfi_1, aes(x=ef_regulation, y=ef_score, color=region))+
  geom_point()+
geom_smooth(method="lm", se=F)
```

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



The scatter plot suggests a strong positive relationship between economic freedom regulation and economic freedom score. For countries in East Asia, the relationship appears to be slightly stronger (better fit) than countries in South Asia.

5. ***(2 pts)*** Compute the correlation coefficient for the variables in exercise 4 above. Comment about the strength of the relationship.

```
new_hfi_1|>
  drop_na(ef_score,ef_regulation)|>
  summarise(cor(ef_regulation,ef_score))
```

```
# A tibble: 1 x 1
  `cor(ef_regulation, ef_score)`
                           <dbl>
1                          0.850
```

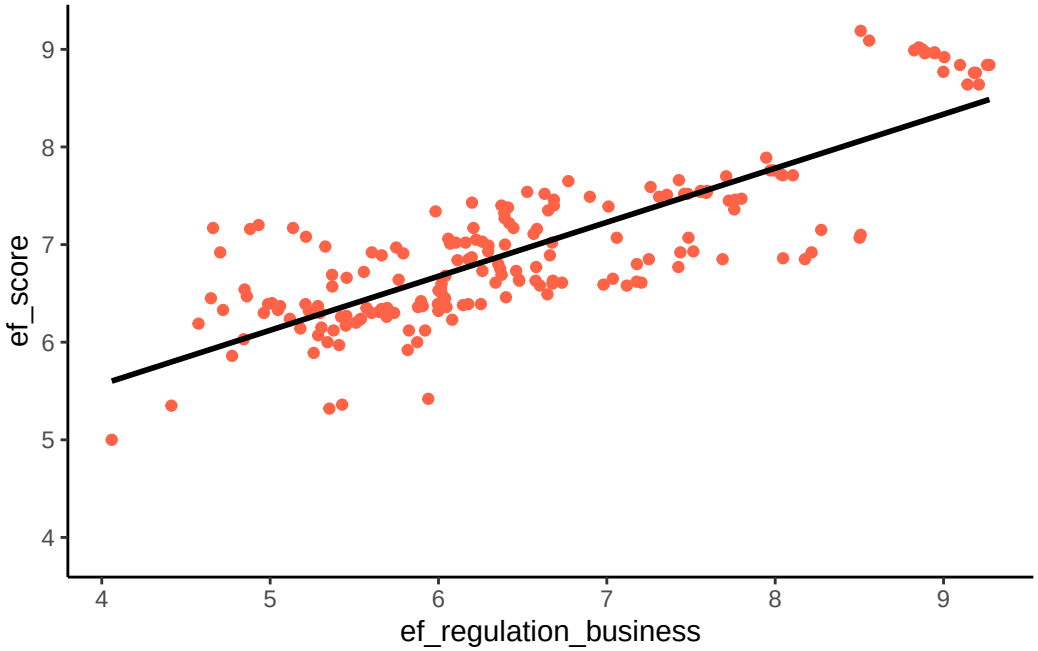
This correlation value is high indicating a linear and strong positive relationship between economic freedom regulation and economic score. Countries with high scores on economic

freedom regulation (i.e., less strict regulations) have a high economic freedom regulation.

6. ***(6 pts)*** Suppose we want to create a model for estimating the economic freedom score (`ef_score`) of countries in East or South Asia based on their business regulation (`ef_regulation_business`). Create a scatter plot to assess the relationship between the variables. Given the scatter plot, do you think it is appropriate to proceed with linear regression modeling? If yes, create the model and interpret its parameters(i.e., the ***slope*** and the ***intercept***) in context. Use your model to estimate the economic freedom score of a country with a business regulation score of 3.9.

```
ggplot(data=new_hfi_1, aes(x=ef_regulation_business,
                           y=ef_score))+
  geom_point(color="tomato")+
  theme_classic()+
  geom_smooth(method="lm", se=F, color="black")
```

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



The scatter plot shows that the relationship between the variables is linear and strong. We can proceed with linear regression modelling.

```
model_2<-lm(ef_score~ef_regulation_business,
            data=new_hfi_1)
tidy(model_2)|>
```

```
  select(term, estimate)
```

```
# A tibble: 2 x 2
  term                    estimate
  <chr>                      <dbl>
1 (Intercept)                3.36
2 ef_regulation_business     0.553
```

7. ***(2 pts)*** Find the R-squared value for the model in exercise 6 and explain its meaning in context.

```
glance(model_2)|>
  select(r.squared)
```

```
# A tibble: 1 x 1
  r.squared
      <dbl>
1     0.678
```

Up to 67.83% of variability in economic freedoms scores can be explained by the model with economic regulation score as the only predictor variable.