Section 4: Simple Linear Regression

In [1]:
```
# RUN THIS CELL
# Load packages
library(testthat)
library(tidyverse) %>% suppressMessages()
```

Load dataset

In [2]:
```
# read in dataset
cho <- read.csv("cho_rep_clean.csv") %>% rename(catholic = catholic2)

# display first 6 rows
head(cho)
```

A dat

	country	htflowsunodc	prostitutionlaw	prostitutionbrothel	ruleWB_m	pop_ln	gdp_pc_const_ppp_ln	de
	<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
1	Aruba	NA	NA	NA	1.6626558	11.29416	11.223195	
2	Andorra	NA	NA	NA	1.5205336	14.62870	10.252641	
3	Angola	0	0	0	-1.5738676	16.34436	7.671666	
4	Albania	3	0	0	-1.2143087	14.95774	8.190376	
5	Netherlands Antilles	NA	NA	NA	0.0336446	12.15826	9.810922	
6	United Arab Emirates	4	0	0	0.9056994	14.70403	10.740041	

The Dataset

How many rows are in the dataset? What does each row represent?

In [3]:
```
n_countries <- nrow(cho)
n_countries
```

```
171
```

Variables

- What is the variable that represents human trafficking severity?
- What is the variable that represents legal prostitution?

Your answer here

Filter to the countries with legalized prostitution.

In [4]:
```
cho_legal <- cho %>% filter(prostitutionlaw == 1)
cho_legal
```

A data.f

country	htflowsunodc	prostitutionlaw	prostitutionbrothel	ruleWB_m	pop_ln	gdp_pc_const_ppp_ln	der
<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
Australia	4	1	1	1.78104460	16.70988	10.147517	
Austria	4	1	0	1.86340666	15.88907	10.224189	
Belgium	5	1	0	1.30211377	16.13168	10.191740	
Bulgaria	3	1	0	-0.24526100	15.94374	8.831015	
Belize	2	1	0	-0.08461525	12.28535	8.484803	
Bolivia	NA	1	0	-0.28332025	15.82830	8.067819	
Brazil	1	1	0	-0.27097645	18.90120	8.952045	
Canada	4	1	0	1.77580798	17.19494	10.232001	
Switzerland	4	1	1	1.99819148	15.76726	10.373944	
Chile	1	1	0	1.15430415	16.48343	9.120388	
Cote d'Ivoire	3	1	0	-0.89846641	16.52228	7.448454	
Costa Rica	2	1	0	0.67662901	15.06215	8.882832	
Cuba	0	1	0	-0.82662666	16.20520	8.833697	
Cyprus	4	1	0	0.82841063	13.50284	9.908668	
Czech Republic	4	1	0	0.84788197	16.15066	9.654954	
Germany	5	1	0	1.68822742	18.21785	10.234450	
Dominican Republic	3	1	0	-0.54071343	15.91029	8.385420	
Ecuador	2	1	1	-0.62846470	16.24972	8.639853	
Spain	4	1	0	1.31303883	17.48895	9.952589	
Estonia	3	1	0	0.51794660	14.17807	8.958530	
Finland	3	1	0	1.95158112	15.44632	9.993948	
France	4	1	0	1.36976433	17.87326	10.135451	
United Kingdom	4	1	0	1.81710422	17.87628	10.141106	
Greece	5	1	0	0.72077054	16.17957	9.785824	
Guatemala	3	1	0	-0.93212140	16.11876	8.206132	
Hong Kong, China	4	1	0	1.07440758	15.63295	10.251199	
Honduras	1	1	0	-0.84103638	15.53613	7.925662	
India	4	1	0	0.23442495	20.65304	7.280837	
Ireland	2	1	0	1.66823471	15.09890	9.976963	
Iceland	3	1	0	1.82918477	12.49874	10.114328	
Israel	5	1	0	1.02552307	15.52841	9.931268	
Italy	5	1	0	0.83276612	17.85586	10.137316	
Kazakhstan	3	1	0	-0.85227269	16.57651	8.411691	
Kiribati	NA	1	0	-0.68682849	11.25492	7.548414	
Lesotho	0	1	0	0.09053291	14.36115	6.917676	

country	htflowsunodc	prostitutionlaw	prostitutionbrothel	ruleWB_m	pop_ln	gdp_pc_const_ppp_ln	der
<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
Luxembourg	2	1	0	1.81542480	12.92055	10.787620	
Latvia	3	1	0	0.20358486	14.73778	8.717217	
Moldova	1	1	0	-0.26917922	15.28310	7.524004	
Mexico	3	1	0	-0.50606650	18.32796	9.207165	
Malaysia	3	1	0	0.47440773	16.84051	9.125359	
Nicaragua	0	1	0	-0.72004890	15.35441	7.507682	
Netherlands	5	1	0	1.79558516	16.55377	10.256337	
Norway	3	1	0	2.00070810	15.28798	10.532573	
New Zealand	3	1	1	1.87833452	15.11663	9.935451	
Panama	3	1	1	-0.16745313	14.79860	8.877958	
Peru	0	1	0	-0.64350736	16.99120	8.574545	
Poland	4	1	0	0.74452591	17.46844	9.104919	
Portugal	3	1	0	1.23328280	16.12079	9.770202	
Paraguay	2	1	0	-0.92719346	15.38451	8.357702	
Senegal	1	1	0	-0.17882787	15.97421	7.222029	
Singapore	3	1	0	1.46912682	15.07525	10.339892	
El Salvador	3	1	0	-0.54845583	15.56094	8.439880	
Slovak Republic	1	1	0	0.22873701	15.49516	9.265897	
Sweden	3	1	0	1.79728246	15.99378	10.111689	
Tonga	NA	1	0	-0.68682849	11.48636	8.173002	
Turkey	5	1	1	-0.07961164	17.92976	9.033379	
Venezuela, RB	3	1	1	-0.71092582	16.90851	9.226381	

Filter to the countries without legalized prostitution.

In [5]:

```r
cho_legal <- cho %>% filter(prostitutionlaw == 0)
cho_legal
```

A data

country	htflowsunodc	prostitutionlaw	prostitutionbrothel	ruleWB_m	pop_ln	gdp_pc_const_ppp_ln	de
<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
Angola	0	0	0	-1.57386756	16.34436	7.671666	
Albania	3	0	0	-1.21430874	14.95774	8.190376	
United Arab Emirates	4	0	0	0.90569937	14.70403	10.740041	
Argentina	3	0	0	0.07710288	17.36432	9.171230	
Armenia	0	0	0	-0.45313463	14.98588	7.441391	
Antigua and Barbuda	NA	0	0	0.99462587	11.12958	9.454256	
Azerbaijan	0	0	0	-1.01244223	15.85478	7.524990	
Bahrain	3	0	0	0.62667704	13.26679	9.962639	
Bahamas, The	NA	0	0	1.20748937	12.54697	10.357580	
Bosnia and Herzegovina	4	0	0	-0.64427644	15.01912	7.399930	
Belarus	0	0	0	-0.75745028	16.13731	8.357536	
Barbados	NA	0	0	1.16897595	12.46068	8.784761	
Brunei Darussalam	2	0	0	0.59473681	12.59460	10.825976	
Botswana	0	0	0	0.55769366	14.25369	8.931231	
China	4	0	0	-0.36122975	20.90963	7.522483	
Cameroon	3	0	0	-1.14665830	16.45845	7.402194	
Colombia	0	0	0	-0.76288164	17.41170	8.824656	
Djibouti	1	0	0	-0.70142704	13.34358	7.653277	
Denmark	4	0	0	1.86016917	15.46954	10.242929	
Algeria	1	0	0	-1.19447780	17.15714	8.636068	
Egypt, Arab Rep.	2	0	0	-0.06246360	17.97217	8.133176	
Fiji	1	0	0	0.05085832	13.55146	8.215861	
Gabon	3	0	0	-0.47349659	13.89658	9.609053	
Georgia	1	0	0	-1.21216977	15.43863	7.455602	
Equatorial Guinea	3	0	0	-1.36790466	13.02141	7.745188	
Guyana	NA	0	0	-0.26927024	13.53956	7.652019	
Croatia	3	0	0	-0.20441842	15.35646	9.202618	
Hungary	3	0	0	0.78287750	16.15046	9.300550	
Indonesia	2	0	0	-0.72285420	19.07041	7.945576	
Iran, Islamic Rep.	3	0	0	-0.44860139	17.89227	8.827672	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Mongolia	NA	0	0	-0.04234968	14.63522	7.536426	
Mauritius	NA	0	0	0.81856507	13.93103	8.896916	
Namibia	NA	0	0	0.29076684	14.29806	8.393890	

country	htflowsunodc	prostitutionlaw	prostitutionbrothel	ruleWB_m	pop_ln	gdp_pc_const_ppp_ln	de
<chr>	<int>	<int>	<int>	<dbl>	<dbl>	<dbl>	
Nigeria	3	0	0	-1.30215168	18.52007	7.254820	
Oman	2	0	0	0.81585735	14.59108	9.727526	
Pakistan	4	0	0	-0.73053432	18.62260	7.526448	
Philippines	3	0	0	-0.08747464	18.06351	7.770010	
Papua New Guinea	NA	0	0	-0.64956945	15.36506	7.719418	
Qatar	3	0	0	0.40734452	13.17255	10.152229	
Romania	2	0	0	-0.13643420	16.93704	8.883523	
Russian Federation	3	0	0	-0.84664845	18.81367	8.968375	
Saudi Arabia	4	0	0	0.31602293	16.72086	9.885262	
Sudan	1	0	0	-1.56829751	17.24436	7.001850	
Serbia	3	0	0	-1.26626420	15.86178	8.648478	
Suriname	NA	0	0	-0.21121214	12.98503	8.507518	
Slovenia	2	0	0	1.10811150	14.50364	9.675705	
Swaziland	0	0	0	-0.58273190	13.78425	8.180374	
Syrian Arab Republic	3	0	0	-0.34821406	16.49724	8.231968	
Thailand	5	0	0	0.48333696	17.91219	8.638437	
Turkmenistan	0	0	0	-1.12494516	15.24759	7.624187	
Trinidad and Tobago	1	0	0	0.36828858	14.05028	9.318598	
Tunisia	0	0	0	-0.01916915	16.00800	8.394404	
Taiwan	NA	0	0	0.86052132	11.00338	9.901899	
Ukraine	3	0	0	-0.92371714	17.75733	8.268387	
Uruguay	0	0	0	0.53253013	14.98433	9.054387	
United States	5	0	0	1.68126619	19.40005	10.431268	
Uzbekistan	2	0	0	-0.97454673	16.94161	7.287391	
Vietnam	3	0	0	-0.45721567	18.10570	7.101788	
Yemen, Rep.	2	0	0	-1.08492434	16.55783	7.521687	
South Africa	3	0	0	0.13574505	17.48215	8.921376	

Simple Linear Regression

A simple linear regression will compare these two groups.

Visual

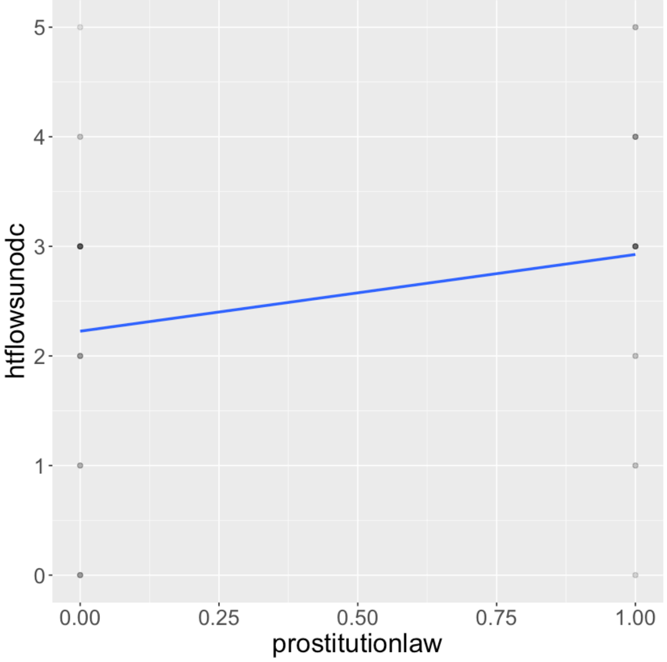
The code below plots a graph with legal status of prostitution on the X axis, and human trafficking severity on the Y axis. The blue line is the line of best fit extracted from the linear regression.

Does it look like there is a significant difference in human trafficking flows between those with legalized prostitution and those without? Is this suggestive that legalizing prostitution could be increase human trafficking?

In [6]:

```
cho %>% ggplot(aes(x =prostitutionlaw, y =  htflowsunodc)) +
geom_point(alpha = 0.05) +
geom_smooth(method = "lm", se = F) +
  theme(text = element_text(size = 20))
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 55 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 55 rows containing missing values (`geom_point()`).”
```



If I have a dataset, can I retrieve the line of best fit between two variables myself?

Yes, you can!

To run the linear regression yourself, all you need to know is the **dependent variable, the independent variable, and the dataset name**. (No, you do not need to know the math to fit the line yourself; r will do it for you! If you are interested in the math, let me know and I can point you at some resources/online videos.)

We use the `lm` function to run a simple linear regression. The `lm` function takes in...

```
lm(dv ~ iv, data = df)
```

It's that simple! So for the cho data, it would be...

In [7]:

```
# EXAMPLE: Bivariate Equation
mod_legal <- lm(htflowsunodc ~ prostitutionlaw, data = cho)

summary(mod_legal)
```

```
Call:
lm(formula = htflowsunodc ~ prostitutionlaw, data = cho)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.92593 -1.00090  0.07407  0.77419  2.77419 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.2258     0.1798  12.378  < 2e-16 ***
prostitutionlaw   0.7001     0.2636   2.656  0.00903 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.416 on 114 degrees of freedom
  (55 observations deleted due to missingness)
Multiple R-squared:  0.05829,	Adjusted R-squared:  0.05003 
F-statistic: 7.057 on 1 and 114 DF,  p-value: 0.009028
```

Interpretation

A one unit increase in legal status of prostituion is associated with a 0.70 increase in the human trafficking index. Correlation, not causation.

Your turn! Let's practice using and interpreting linear regression with two variables

Note: This is real data. So the relationships your observing are reflective of the real world. Isn't that cool?

Q1) Democracy and Human Trafficking?

Run a regression where we investigate: Does democracy have a relationship with human trafficking flows?

- IV: `democracy`
- DV: `htflowsunodc`

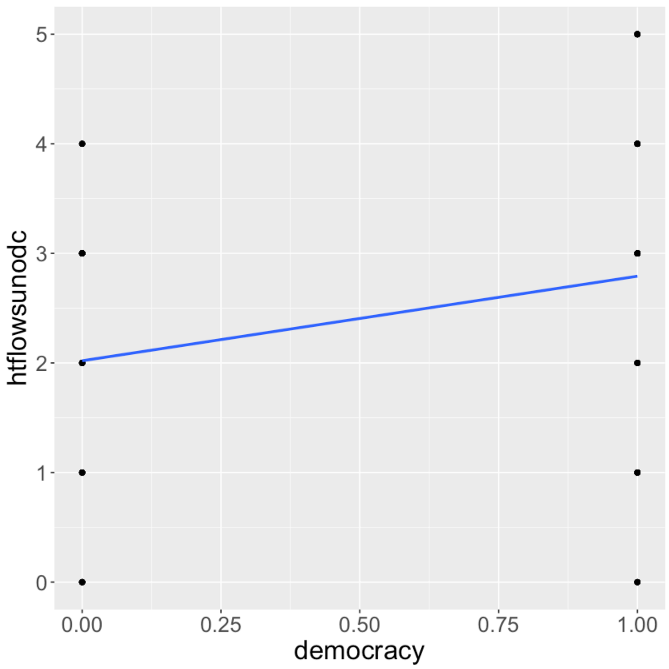
$$htflowsunodc = \alpha + \beta_1 democracy + \epsilon_i$$

Visually, it looks like there is a relationship.

In [8]:

```
# RUN, No need to edit
cho %>% ggplot(aes(x =democracy, y =  htflowsunodc)) +
geom_point() +
geom_smooth(method = "lm", se = F) +
  theme(text = element_text(size = 20))
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 49 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 49 rows containing missing values (`geom_point()`).”
```



Your task: Use `lm` to estimate the relationship.

In [9]:

```
# YOUR ANSWER HERE
mod1 <- lm(htflowsunodc~democracy, data = cho) # YOUR CODE HERE

summary(mod1)
```

```
Call:
lm(formula = htflowsunodc ~ democracy, data = cho)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.7917 -1.0200  0.2083  0.9800  2.2083 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.0200     0.2004  10.082  < 2e-16 ***
democracy     0.7717     0.2608   2.959  0.00372 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.417 on 120 degrees of freedom
  (49 observations deleted due to missingness)
Multiple R-squared:  0.06799,	Adjusted R-squared:  0.06023 
F-statistic: 8.754 on 1 and 120 DF,  p-value: 0.003722
```

In [10]:

```
. = ottr::check("tests/Q1.R")
```

```
Test Q1 - 1 passed
```

Interpretation

How would you interpret $\alpha$ and $\beta_1$? Your value for $\alpha$ should be 2.02, and $\beta_1$ should be 0.77. Are democracies associated with higher rates of trafficking inflow?

Non democracies have an average score of 2.02 on the human trafficking index. Democracies are associated with a 0.77 points increase, on average, in the human trafficking flows index. In other words, democracies have an average score of 2.79 on the human trafficking index.

Yes.

Q2) Democracy and Legalized Prostitution?

Run a regression where we investigate: Does gdp per capita have a relationship with human trafficking flows?

- IV: `democracy`
- DV: `prostitutionlaw`

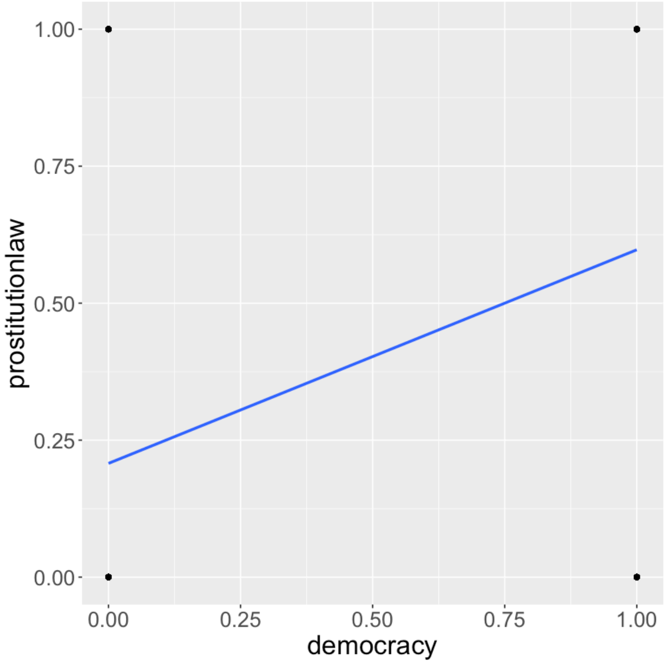
$$prostitutionlaw = \alpha + \beta_1 democracy + \epsilon_i$$

Visually, it looks like there is a relationship.

In [11]:

```
# RUN, No need to edit
cho %>% ggplot(aes(x =democracy, y =  prostitutionlaw)) +
geom_point() +
geom_smooth(method = "lm", se = F) +
  theme(text = element_text(size = 20))
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 41 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 41 rows containing missing values (`geom_point()`).”
```



Your task: Use `lm` to estimate the relationship.

In [12]:

```
# YOUR ANSWER HERE
mod2 <- lm(prostitutionlaw ~ democracy, data = cho) # YOUR CODE HERE

summary(mod2)
```

```
Call:
lm(formula = prostitutionlaw ~ democracy, data = cho)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.5974 -0.2076 -0.2076  0.4026  0.7924 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.20755    0.06336   3.276  0.00136 ** 
democracy    0.38986    0.08233   4.735 5.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4613 on 128 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.1491,	Adjusted R-squared:  0.1424 
F-statistic: 22.42 on 1 and 128 DF,  p-value: 5.712e-06
```

```
In [13]: . = ottr::check("tests/Q2.R")
```

```
Test Q2 - 1 passed
```

Interpretation

There is a relationship!

Because prostitution law is a binary variable (takes on 0/1 values), you can interpret this as:

- $\alpha = 0.20755$: Non-democracies have an average probability of 0.21 for legalizing prostitution.
- $\beta_1 = 0.3899$: On average, relative to non-democracies, democracies are associated with 0.39 higher probability of legalized prostitution.

Are democracies, on average, associated with a higher probability of legalizing prostitution? Which coefficient tells you this–α or β_1? What if β_1 were negative?

Replace this text

Q3) Wait a minute...Could democracy be driving this relationship?

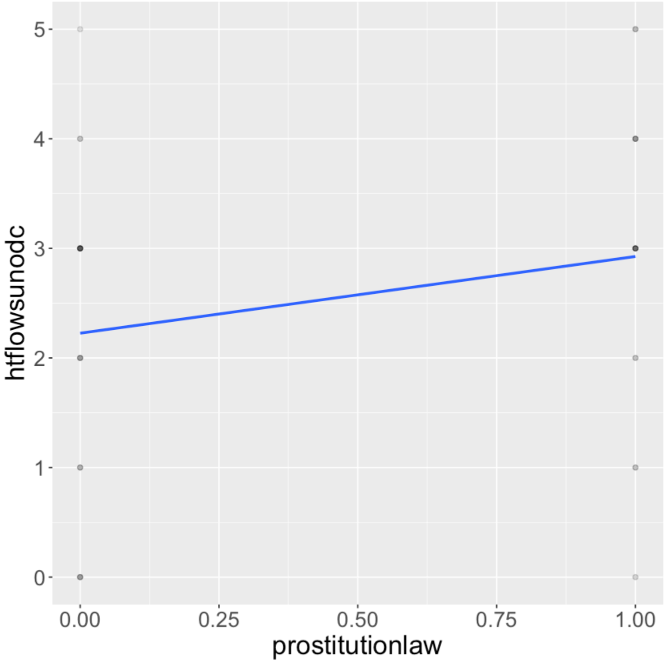
In other words, we could theorize democracies are more likely to legalize prostitution and democracies are likely to be recording human trafficking inflows.

Is the relationship that we see (between legalized prostitution and human trafficking) simply be because democracies are more proactively documenting human trafficking? Legalizing prostitution doesn't actually make a difference and democracy is the alternative explanation of why we see this relationship.

More on this next time.

```
In [14]: cho %>% ggplot(aes(x =prostitutionlaw, y =  htflowsunodc)) +
         geom_point(alpha = 0.05) +
         geom_smooth(method = "lm", se = F) +
           theme(text = element_text(size = 20))
```

```
`geom_smooth()` using formula = 'y ~ x'
Warning message:
“Removed 55 rows containing non-finite values (`stat_smooth()`).”
Warning message:
“Removed 55 rows containing missing values (`geom_point()`).”
```



In your own words, why is democracy an alternative explanation to this relationship?

Your answer here

Extra Time

Explore more relationships in the code chunk below using `the` lm function:

1. Are countries with higher gdp per capita (`gdp_pc_const_ppp_ln`) associated with lower rates of trafficking (`htflowsunodc`)? What is the estimated relationship?
2. Are countries with higher shares of catholics (`catholic`) associated with lower rates of legalization (`prostitutionlaw`)? What is the estimated relationship?
3. Are countries in West Europe (`reg_west_europe`) associated with higher rates of legalization (`prostitutionlaw`) relative to the rest of the countries?
4. What about sub-saharan africa (`reg_ssa`)?
5. What about latin america (`reg_latam`)?

In [ ]:

```
# YOUR CODE HERE
```

Preview: Accounting for Democracy

The cell below is the original regression evaluateing the association between legalizing prostituion and human trafficking.

```
lm(dv ~ iv, data = df)
```

In [15]:

```
# RUN CELL DO NOT CHANGE
lm(htflowsunodc ~ prostitutionlaw, data = cho) %>% summary()
```

```
Call:
lm(formula = htflowsunodc ~ prostitutionlaw, data = cho)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.92593 -1.00090  0.07407  0.77419  2.77419 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       2.2258     0.1798  12.378  < 2e-16 ***
prostitutionlaw   0.7001     0.2636   2.656  0.00903 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.416 on 114 degrees of freedom
  (55 observations deleted due to missingness)
Multiple R-squared:  0.05829,	Adjusted R-squared:  0.05003 
F-statistic: 7.057 on 1 and 114 DF,  p-value: 0.009028
```

The cell below accounts for democracy as a potential alternative explanation. We add in democracy on the right hand side as a "control" variable.

The cell below reads: What is the association of prostitutionlaw on human trafficking flows, holding democracy constant.

Example code:

```
lm(dv ~ iv + control, data = df)
```

In [16]:

```
# RUN CELL DO NOT CHANGE
lm(htflowsunodc ~ prostitutionlaw + democracy, data = cho) %>% summary()
```

```
Call:
lm(formula = htflowsunodc ~ prostitutionlaw + democracy, data = cho)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.04615 -1.04615  0.02551  1.02551  2.37628 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.9745     0.2096   9.419 6.75e-16 ***
prostitutionlaw   0.4224     0.2874   1.470   0.1444    
democracy         0.6492     0.2911   2.230   0.0277 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.392 on 113 degrees of freedom
  (55 observations deleted due to missingness)
Multiple R-squared:  0.09799,	Adjusted R-squared:  0.08203 
F-statistic: 6.138 on 2 and 113 DF,  p-value: 0.002947
```

Look as the value for `prostitutionlaw` in the `Estimate` column for both regression outputs above. This represents $\beta_1$. How does this change before and after including democracy? Why?

Your answer here

Summary

We are using linear regression (via the `lm` function) to quantify the relationship between X and Y. As of right now, this is purely a correlational relationship, not a causal relationship.

This relationship is written as: $$Y= \alpha + \beta_1 X + \epsilon_i$$

α is interpreted as: The value of Y when X = 0.
β_1 is interpreted as: A one unit increase in X is associated with a β_1 unit increase in Y.

Next Time

Since we can theorize these alternative explanations to this relationship, is there a way we can account for these in the regression?

Yes:) We can try to isolate for the effect of legalizing prostitution by adding "controls", more next time!

What you need to know for Assignment #2

Already covered:

- Direction of Correlation
- Independent variable, dependent variables
- How do you interpret $Y = \alpha + \beta X$

Next time:

- “Control” variables
- lm() with control variables