Linear regression review

Frank Edwards

February 10, 2021

Linear regression: IPV data

```
ipv<-read_csv("./slides/data/dhs_ipv.csv")

head(ipv)
```

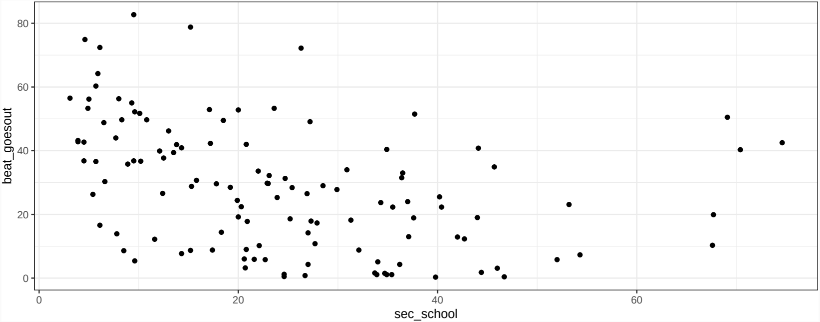
```
## # A tibble: 6 x 7
##   beat_burnfood beat_goesout sec_school no_media country    year region
##           <dbl>        <dbl>      <dbl>    <dbl> <chr>     <dbl> <chr>
## 1           4.4         18.6       25.2      1.5 Albania    2008 Middle East an~
## 2           4.9         19.9       67.7      8.7 Armenia    2000 Middle East an~
## 3           2.1         10.3       67.6      2.2 Armenia    2005 Middle East an~
## 4           0.3          3.1       46        6.4 Armenia    2010 Middle East an~
## 5          12.1         42.5       74.6      7.4 Azerbaij~  2006 Middle East an~
## 6          NA           NA         24       41.9 Banglade~  2004 Asia
```

Research question

- Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?

Visualizing associations: scatterplots

```
ggplot(ipv,
       aes(x = sec_school, y = beat_goesout)) +
  geom_point()
```

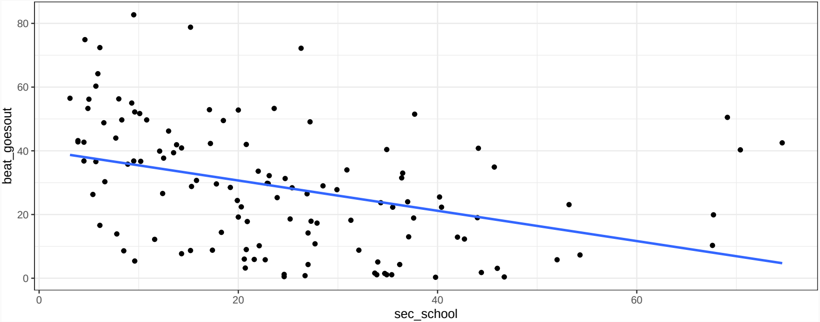


Describing linear associations: correlation

```
cor(ipv$sec_school, ipv$beat_goesout, use = "complete")
```

```
## [1] -0.3802336
```

```
ggplot(ipv,
       aes(x = sec_school, y = beat_goesout)) +
  geom_point() +
  geom_smooth(method = "lm", se = F)
```



The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y. Allows us to estimate the line, even when x and y do not fall exactly on a line.

The linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

β_0: The value of y when x is equal to zero

β_1: The average increase in y when x increases by one unit

ε: The distance between the line $y = \beta_0 + \beta_1 X$ and the actual observed values of y. Allows us to estimate the line, even when x and y do not fall exactly on a line.

The line $y = \beta_0 + \beta_1 X$ provides a prediction for the values of y based on the values of x.

Understanding the regression line for real data

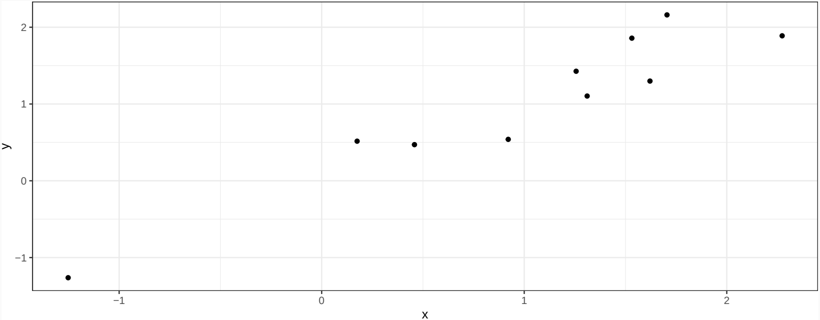
```
## # A tibble: 10 x 2
##         x      y
##     <dbl>  <dbl>
##  1  0.458  0.471
##  2 -1.25  -1.26
##  3  1.26   1.43
##  4  1.53   1.86
##  5  2.27   1.89
##  6  1.62   1.30
##  7  0.921  0.540
##  8  1.70   2.16
##  9  0.175  0.516
## 10  1.31   1.10
```

$\beta_0 = 0.05$, $\beta_1 = 0.95$

- Estimate $\hat{Y}$. Recall that $\hat{Y} = \beta_0 + \beta_1 X$
- Estimate ε. Recall that $\varepsilon = Y - \hat{Y}$

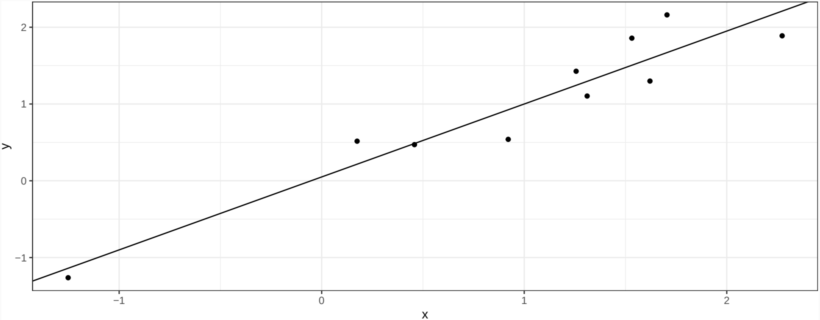
Understanding the regression line

$\beta_0 = 0.05$, $\beta_1 = 0.95$



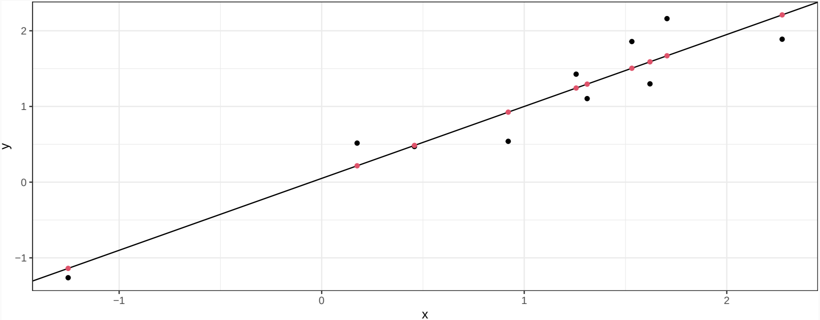
Understanding the regression line: adding the fit

$\beta_0 = 0.05$, $\beta_1 = 0.95$



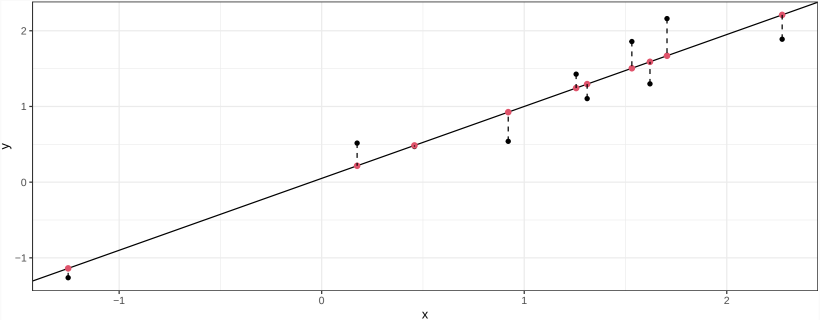
Understanding the regression line: adding $\hat{y}$

$\beta_0 = 0.05$, $\beta_1 = 0.95$



Understanding the regression line: adding ε

$\beta_0 = 0.05$, $\beta_1 = 0.95$



Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?

Write a linear regression formula that will allow us to test this question.

Asking a question with regression

Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?

Write a linear regression formula that will allow us to test this question.

Two ways of writing a linear regression

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

Two ways of writing a linear regression

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2)$$

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1$$

Estimating a regression model in R

```r
library(broom)
## models take the general form
## lm(outcome ~ predictor, data)
ipv_model<-lm(beat_goesout ~ sec_school,
              data = ipv)

tidy(ipv_model)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   40.2       3.07      13.1  6.98e-25
## 2 sec_school    -0.475     0.106     -4.50 1.56e- 5
```

Interpret this model

Using rstanarm

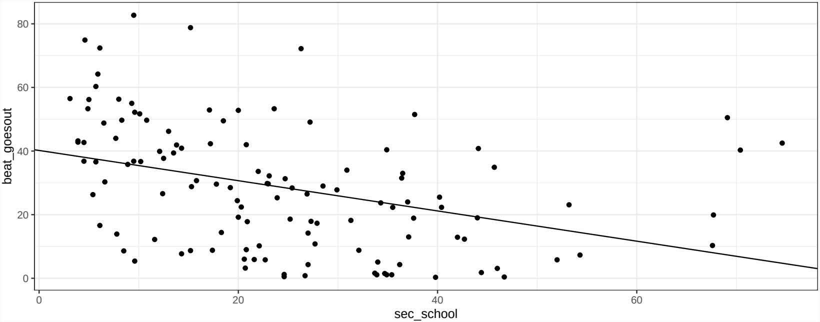
```r
library(rstanarm)
library(broom.mixed)

ipv_model_stan<-stan_glm(beat_goesout ~ sec_school,
                         data = ipv)
```

```
##
## SAMPLING FOR MODEL 'continuous' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 8.3e-05 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 2000 [  0%]  (Warmup)
## Chain 1: Iteration:  200 / 2000 [ 10%]  (Warmup)
## Chain 1: Iteration:  400 / 2000 [ 20%]  (Warmup)
```

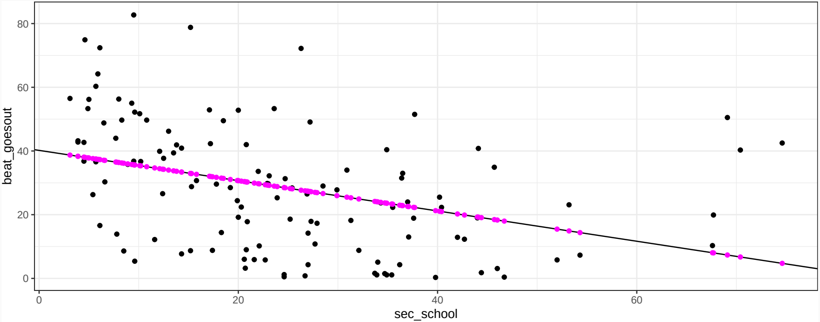
Visualize the model

```
ggplot(ipv %>%
         filter(!(is.na(sec_school)), !(is.na(beat_goesout))),
       aes(x=sec_school, y = beat_goesout)) +
  geom_point() +
  geom_abline(aes(intercept = coef(ipv_model)[1],
                  slope = coef(ipv_model)[2]))
```



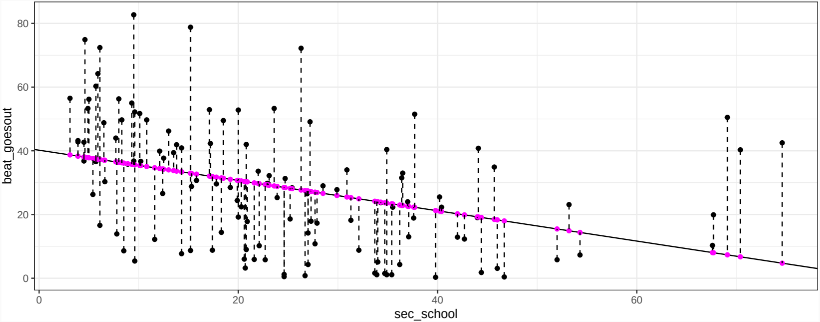
Visualize the model: expected values of y

```
ggplot(ipv %>%
         filter(!(is.na(sec_school)), !(is.na(beat_goesout))),
       aes(x=sec_school, y = beat_goesout)) +
  geom_point() +
  geom_abline(aes(intercept = coef(ipv_model)[1],
                  slope = coef(ipv_model)[2])) +
  geom_point(aes(x = sec_school, y = fitted(ipv_model)), color = "magenta")
```



Visualize the model: error term (residuals)

```
ggplot(ipv %>%
         filter(!(is.na(sec_school)), !(is.na(beat_goesout))),
       aes(x=sec_school, y = beat_goesout)) +
  geom_point() +
  geom_abline(aes(intercept = coef(ipv_model)[1],
                  slope = coef(ipv_model)[2])) +
  geom_point(aes(x = sec_school, y = fitted(ipv_model)), color = "magenta") +
  geom_segment(aes(x = sec_school, xend = sec_school,
                   y = beat_goesout, yend = fitted(ipv_model)), lty =2)
```



Linear regression with multiple predictors

We can extend the linear regression model:

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1$$

to include more than one predictor.

Linear regression with multiple predictors

We can extend the linear regression model:

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1$$

to include more than one predictor.

We rewrite the equation as:

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots \beta_p x_p$$

Linear regression with multiple predictors

We can extend the linear regression model:

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1$$

to include more than one predictor.

We rewrite the equation as:

$$y \sim N(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots \beta_p x_p$$

In matrix notation

To be more compact:

$$Y = \beta X + \varepsilon$$

Where Y is the vector of predictors, β is the vector of coefficients (including the intercept), X is the matrix of all predictors, and ε is the error term.

Linear regression with multiple predictors

Let's start with a single predictor for region

```
m2<-lm(beat_goesout ~ region,
               data = ipv)
```

Interpreting a regression model with intercepts only

```r
coef(m2)
```

```
##                         (Intercept)                regionLatin America
##                           18.673684                         -11.628684
## regionMiddle East and Central Asia           regionSub-Saharan Africa
##                            7.501316                          19.465446
```

```r
ipv %>% group_by(region) %>%
  summarise(beat_goesout = mean(beat_goesout, na.rm=T))
```

```
## # A tibble: 4 x 2
##   region                           beat_goesout
## * <chr>                                   <dbl>
## 1 Asia                                    18.7
## 2 Latin America                            7.04
## 3 Middle East and Central Asia            26.2
## 4 Sub-Saharan Africa                      38.1
```

Add a continuous predictor

```
m3<-lm(beat_goesout~sec_school + region,
       data = ipv)

coef(m3)
```

```
##                      (Intercept)                        sec_school
##                       27.9790347                        -0.3317727
##              regionLatin America regionMiddle East and Central Asia
##                      -11.2761321                        13.7311661
##         regionSub-Saharan Africa
##                       15.8675474
```

Recall that $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

What do we predict will be the level of tolerance for IPV among women

- if sec_school = 50 and region = Latin America

Add a continuous predictor

```r
m3<-lm(beat_goesout~sec_school + region,
       data = ipv)

coef(m3)
```

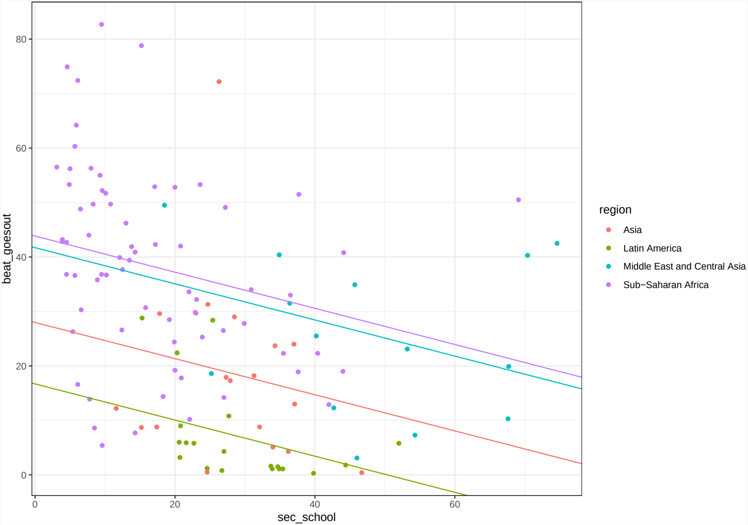
```
##                        (Intercept)                          sec_school
##                         27.9790347                          -0.3317727
##                regionLatin America regionMiddle East and Central Asia
##                        -11.2761321                          13.7311661
##           regionSub-Saharan Africa
##                         15.8675474
```

Recall that $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

What do we predict will be the level of tolerance for IPV among women

- if sec_school = 50 and region = Latin America
- if sec_school = 50 and region = Middle East and Central Asia

Visualizing the model



Interactions

The prior model allowed each region to have its own starting level of tolerance for IPV. What if we thought the relationship (effect) of secondary schooling on IPV depended on region?

We can add *interaction terms* to our model to model processes where we believe the relationship between y and x_1 is a function of x_2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Estimating interactions in R

```r
ipv_model3<-lm(beat_goesout ~ sec_school + region +
                 region * sec_school,
               data = ipv)
```

Interpreting an interaction model

```r
coef(ipv_model3)
```

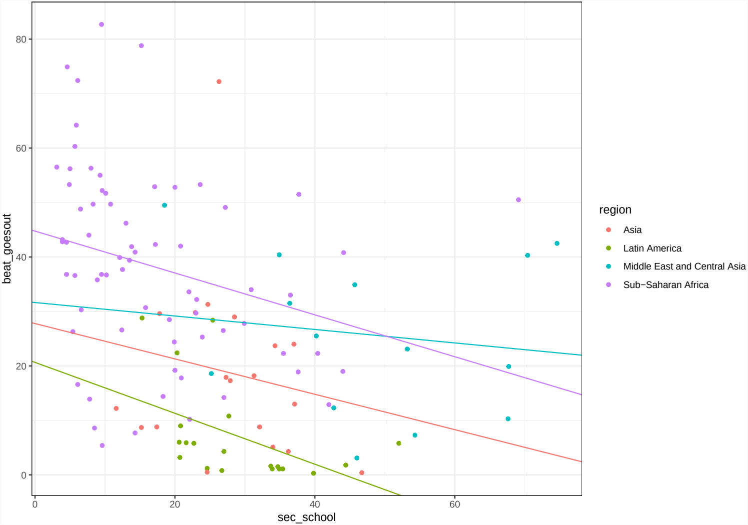
```
##                                    (Intercept)
##                                    27.78048328
##                                     sec_school
##                                    -0.32469353
##                            regionLatin America
##                                    -7.13303634
##          regionMiddle East and Central Asia
##                                     3.85875152
##                      regionSub-Saharan Africa
##                                    16.97257959
##                 sec_school:regionLatin America
```

How interactions work

- What is the predicted level of IPV tolerance in a country where sec_school = 20 in Latin America?
- In Sub-Saharan Africa?

Recall that Asia is the reference category

Visualizing interactions



Practice with real data

```
budget<-read_csv("./slides/data/revenue_dat.csv")

glimpse(budget)
```

```
## Rows: 4,286
## Columns: 33
## $ year_range        <chr> "2007-2011", "2007-2011", "2007-2011", "2007-2011...
## $ fips_st           <chr> "01", "01", "01", "01", "01", "01", "01", "01", "...
## $ fips_cnty         <chr> "001", "005", "007", "009", "011", "013", "015", ...
## $ deaths            <dbl> 3, 1, 0, 0, 0, 1, 5, 1, 0, 1, 1, 1, 0, 1, 0, 0, 0...
## $ exp_tot           <dbl> 49742600, 28588200, 13036120, 36644480, 10940520,...
## $ exp_correction    <dbl> 2101800, 1037880, 80600, 1703760, 0, 487320, 3881...
## $ exp_police        <dbl> 9306200, 5537840, 2421720, 6853480, 2285320, 4067...
## $ exp_welfare       <dbl> 636120, 29760, 2480, 168640, 297600, 358360, 9969...
## $ rev_tot           <dbl> 56454720, 33706920, 13601560, 33338640, 11783720,...
## $ rev_fines         <dbl> 538160, 617520, 124000, 652240, 64480, 111600, 15...
## $ rev_gen_ownsource <dbl> 46527280, 24709480, 8257160, 23612080, 6335160, 1...
## $ rev_int_gov       <dbl> 9626120, 5992920, 5344400, 7676840, 5448560, 6099...
## $ rev_prop_tax      <dbl> 7847960, 5124920, 1883560, 6288040, 2630040, 3842...
## $ rev_tax           <dbl> 34602200, 18292480, 6363680, 17807640, 4615280, 1...
## $ pop_tot           <dbl> 53944, 27546, 22746, 57140, 10877, 20860, 117614,...
## $ pop_pct_men_15_34 <dbl> 0.12742844, 0.15889784, 0.12872593, 0.12779139, 0...
## $ pop_wht           <dbl> 41653, 12941, 17084, 50891, 2431, 11338, 86884, 2...
## $ pop_blk           <dbl> 9755, 12632, 5153, 806, 7619, 9091, 24103, 1283, ...
## $ pop_ami           <dbl> 114, 92, 77, 342, 18, 30, 330, 181, 253, 0, 492, ...
## $ pop_api           <dbl> 385, 147, 12, 48, 18, 164, 844, 60, 103, 45, 107,...
## $ pop_lat           <dbl> 1298, 1344, 375, 4475, 684, 91, 3720, 331, 3052, ...
## $ pop_pct_pov       <dbl> 0.1087869, 0.2471759, 0.1565489, 0.1370827, 0.259...
## $ pop_pct_deep_pov  <dbl> 0.05094139, 0.11641609, 0.06211180, 0.05244872, 0...
```

Let's build a theory: police spending

What variables might be associated with police spending across places?

```
names(budget)
```

```
##  [1] "year_range"        "fips_st"           "fips_cnty"
##  [4] "deaths"            "exp_tot"           "exp_correction"
##  [7] "exp_police"        "exp_welfare"       "rev_tot"
## [10] "rev_fines"         "rev_gen_ownsource" "rev_int_gov"
## [13] "rev_prop_tax"      "rev_tax"           "pop_tot"
## [16] "pop_pct_men_15_34" "pop_wht"           "pop_blk"
## [19] "pop_ami"           "pop_api"           "pop_lat"
## [22] "pop_pct_pov"       "pop_pct_deep_pov"  "pop_med_income"
## [25] "pop_pc_income"     "violent.yr"        "property.yr"
## [28] "murder.yr"         "ft_sworn"          "cbsa"
## [31] "metroname"         "dissim_bw"         "dissim_wl"
```

Describe a linear model that matches the concepts we developed in our theory

Let’s estimate the model

Fit the model using `lm()`

Interpret the model

What is the meaning of each β parameter? What is the meaning of the standard deviation of this estimate?

Visualize the model

Construct an appropriate visual for this model to aid in interpretation

Add an appropriate interaction

Theorize then model an interaction that makes sense for this model

Visualize the interaction

Visualize the interaction for multiple values of each predictor

Interpret your model

Make sure you interpret the model in terms of *conditional means*. Do *not* interpret this model using causal language. We haven't designed for causal inference, but can use the model descriptively.

Repeat this process with a new outcome: `violent.yr`. This variable is a measure of the average number of violent index crimes known to police in a county per year. Build a model that helps us explain variation in violent crime across counties and/or over time.

1. Describe your theory in plain english.
2. Describe a linear model that matches your theory (provide equations for the model).
3. Fit that model using `lm()`
4. Interpret the parameter estimates for that model.
5. Visualize this model
6. Add an appropriate interaction term. Explain your choice.
7. Visualize this new model with an interaction
8. Interpret the model