Linear Regression

Frank Edwards

11/9/21

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between $\hat{Y}$ and Y.

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between $\hat{Y}$ and Y.
- To do so, we minimize the sum of squared residuals (SSR)

Ordinary least squares

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between $\hat{Y}$ and Y.
- To do so, we minimize the sum of squared residuals (SSR)

In other words, we solve for the values of β_0 and β_1 that results in the smallest possible value for:

$$\mathrm{SSR} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X)^2$$

We usually fit a linear regression using a method called *ordinary least squares*, or OLS.

- This method seeks to minimize the distance between $\hat{Y}$ and Y.
- To do so, we minimize the sum of squared residuals (SSR)

In other words, we solve for the values of β_0 and β_1 that results in the smallest possible value for:

$$\mathrm{SSR} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 X)^2$$

Also note that we can estimate the coefficient vector β_1 using matrix algebra:

$$\beta = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y$$

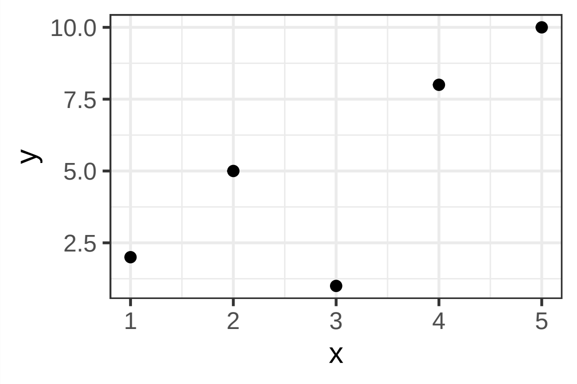
Estimating a regression model in R, the basics

```
x<-c(1, 2, 3, 4, 5)
y<-c(2, 5, 1, 8, 10)

model_demo<-lm(y~x)

coef(model_demo)
```

```
## (Intercept)           x
##        -0.5         1.9
```

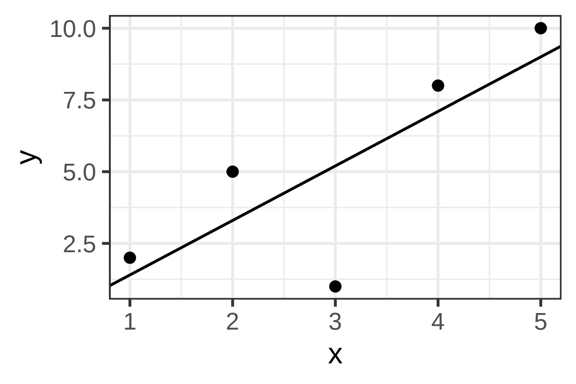
The observed data

```r
ggplot(data.frame(x = x, y = y),
       aes(x = x, y = y)) +
  geom_point()
```



The regression line

```
ggplot(data.frame(x = x, y = y),
       aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept = coef(model_demo)[1],
              slope = coef(model_demo)[2])
```



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

Asking a question with regression

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

- What is the outcome variable (y)? What is the predictor variable (x)?

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

- What is the outcome variable (y)? What is the predictor variable (x)?

E[acceptance of intimate partner violence] = Intercept + Slope $\times$ secondary school completion

Asking a question with regression

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

- What is the outcome variable (y)? What is the predictor variable (x)?

E[acceptance of intimate partner violence] = Intercept + Slope $\times$ secondary school completion

- What is our implied prediction about the slope?

Estimating a regression model in R

```
ipv<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/dhs_ipv.csv")
## models take the general form
## lm(outcome ~ predictor, data)
ipv_model<-lm(beat_goesout ~ sec_school,
              data = ipv)

coef(ipv_model)
```

```
## (Intercept)  sec_school
##  40.1876597  -0.4753799
```

- What does the intercept coefficient (β_0) indicate?

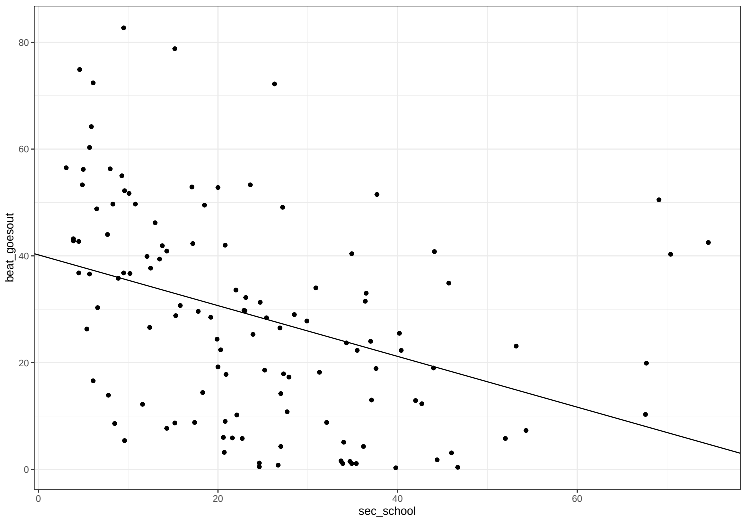
Estimating a regression model in R

```
ipv<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/dhs_ipv.csv")
## models take the general form
## lm(outcome ~ predictor, data)
ipv_model<-lm(beat_goesout ~ sec_school,
              data = ipv)

coef(ipv_model)
```

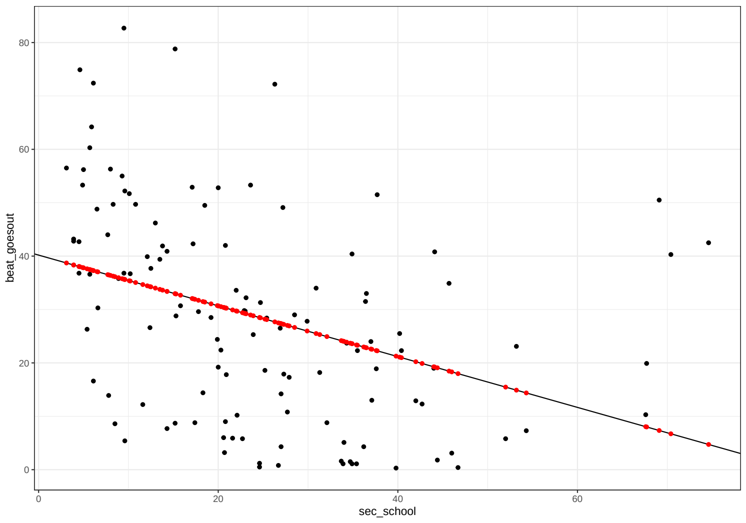
```
## (Intercept)  sec_school
##  40.1876597  -0.4753799
```

- What does the intercept coefficient (β_0) indicate?
- What does the slope coefficient (β_1) indicate?

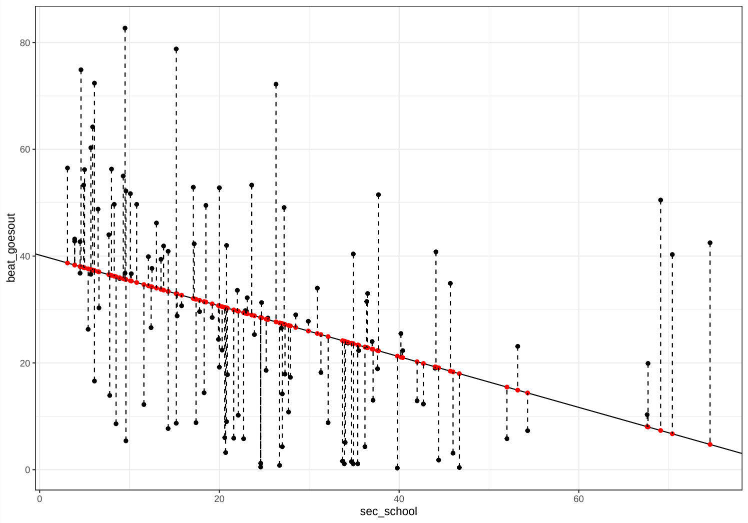
Visualize the model



Visualize the model: expected values of y



Visualize the model: error term (residuals)



Interpreting a regression model

```
coef(ipv_model)
```

```
## (Intercept)  sec_school
##  40.1876597  -0.4753799
```

On average, women in countries where women have higher levels of secondary education have lower levels of acceptance of domestic violence. For example, the model predicts that $\hat{y} = \beta_0 = 40.19$ percent of women in a country in which zero percent of women have a secondary education approve of a husband beating a wife if she goes out without telling him. In a country where 20 percent of women have a secondary education, by contrast, this model predicts that $\hat{y} = \beta_0 + \beta_1 \times 20 = 30.68$ percent of women approve of intimate partner violence for a women going out without notifying her husband, a clear decline. There is a negative linear relationship between average levels of secondary schooling and women's attitudes about intimate partner violence across countries.

We can extend the linear regression model:

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

Linear regression with multiple predictors

We can extend the linear regression model:

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

to include more than one predictor. We rewrite the equation as:

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

Linear regression with multiple predictors

We can extend the linear regression model:

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

to include more than one predictor. We rewrite the equation as:

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

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

Our first model, for country i, was:

$$\text{IPV Attitudes}_i = \beta_0 + \beta_1 \text{Secondary School}_i + \varepsilon$$

Linear regression with multiple predictors

Our first model, for country *i*, was:

$$\text{IPV Attitudes}_i = \beta_0 + \beta_1 \text{Secondary School}_i + \varepsilon$$

Let's add a predictor for region. Remember from prior examples that we saw clear patterns within regions.

Estimating a multiple linear regression in R

```r
ipv_model2<-lm(beat_goesout ~ sec_school + region,
               data = ipv)
```

Interpreting a regression model with multiple coefficients

```r
coef(ipv_model2)
```

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

Now, we have a coefficient for secondary school, in addition to a coefficient for each region. Note that this kind of model requires a “reference category”, which is left out. In this case, Asia is the reference.

Interpreting a regression model with multiple coefficients

```
coef(ipv_model2)
```

```
##                     (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

Interpreting a regression model with multiple coefficients

```r
coef(ipv_model2)
```

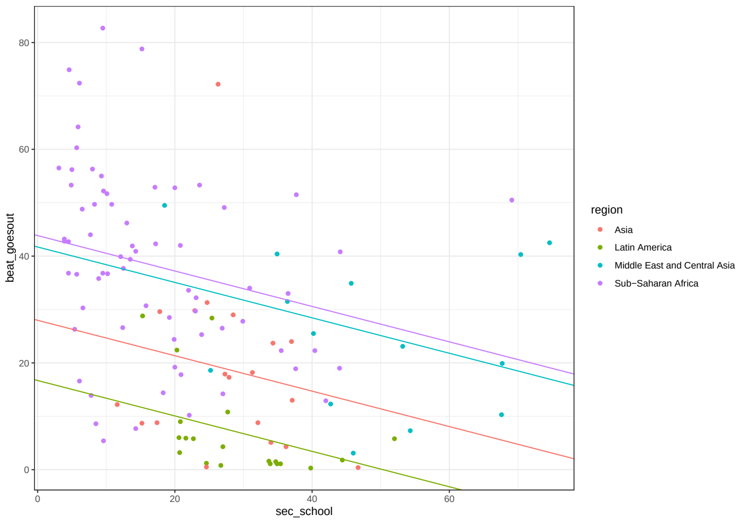
```
##                    (Intercept)                           sec_school
##                     27.9790347                           -0.3317727
##            regionLatin America regionMiddle East and Central Asia
##                    -11.2761321                           13.7311661
##          regionSub-Saharan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

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

Interpreting an interaction model

```
coef(ipv_model3)
```

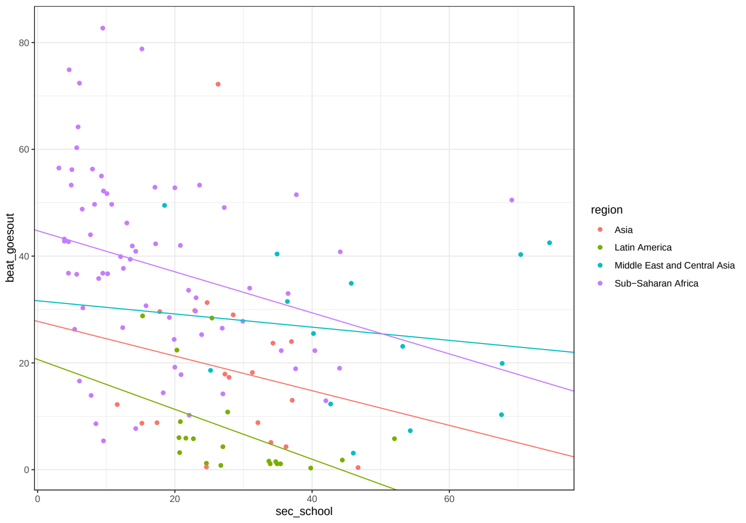
```
##                                (Intercept)
##                                27.78048328
##                                 sec_school
##                                -0.32469353
##                        regionLatin America
##                                -7.13303634
##       regionMiddle 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



Let’s try this with different data

```r
### load in the 'mtcars' data
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3
```

Variables in the mtcars data

- mpg Miles/(US) gallon
- cyl Number of cylinders
- disp Displacement (cu.in.)
- hp Gross horsepower
- drat Rear axle ratio
- wt Weight (1000 lbs)
- qsec 1/4 mile time
- vs Engine (0 = V-shaped, 1 = straight)
- am Transmission (0 = automatic, 1 = manual)
- gear Number of forward gears
- carb Number of carburetors

Let's build a model for fuel efficiency with this data (theory!)

Our outcome of interest is `mpg`. What measured features of these cars do we think might be related to fuel efficiency?

Start simple

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i$$

Start simple

$$\mathrm{E}[\mathrm{mpg}_i] = \beta_0 + \beta_1 \mathrm{hp}_i$$

$$\mathrm{mpg}_i = \beta_0 + \beta_1 \mathrm{hp}_i + \varepsilon$$

How would we estimate this model in R?

Lab: Improve model fit for mpg

How can we make a better model?

Lab part 2: transforming predictors

Load the gapminder data

```
library(gapminder)
```

Let's build a model for life expectancy.

Load the gapminder data

```
library(gapminder)
```

Let's build a model for life expectancy.

$$\text{lifeExp}_i = ??$$

What is life expectancy a function of?

Lab part 3: Prediction

- using `predict()`
- Visualizing model inferences

Conclusion

- Regression models are at the core of social science methodology. Get comfortable with them.
- All models are wrong, some are useful. Reality is rarely accurately described by straight lines, but we can learn a lot from them.
- Think carefully about your modeling decisions. Connect your models to your theory about how a process works.