Prediction, 2

Frank Edwards

December 09, 2019

Linear regression: IPV data

```
ipv <- read_csv("./data/dhs_ipv.csv") %>% select(-X1)

head(ipv)
```

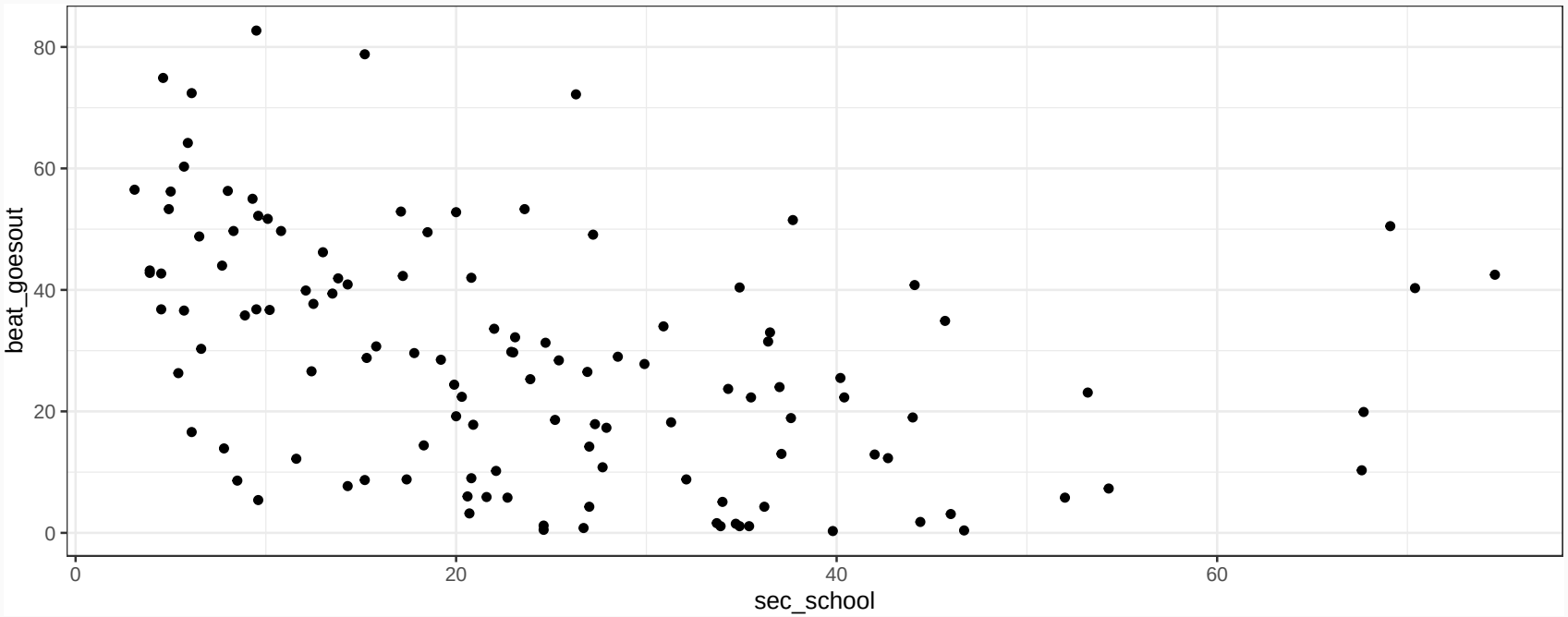
```
## # A tibble: 6 x 7
##   beat_burnfood beat_goesout sec_school no_media country   year region
##           <dbl>        <dbl>      <dbl>    <dbl> <chr>    <dbl> <chr>
## 1           4.4         18.6       25.2      1.5 Albania   2008 Middle Eas~
## 2           4.9         19.9       67.7      8.7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

Research question

- Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?
- I expect a negative association between schooling and tolerance for intimate partner violence.

Visualizing associations: scatterplots

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

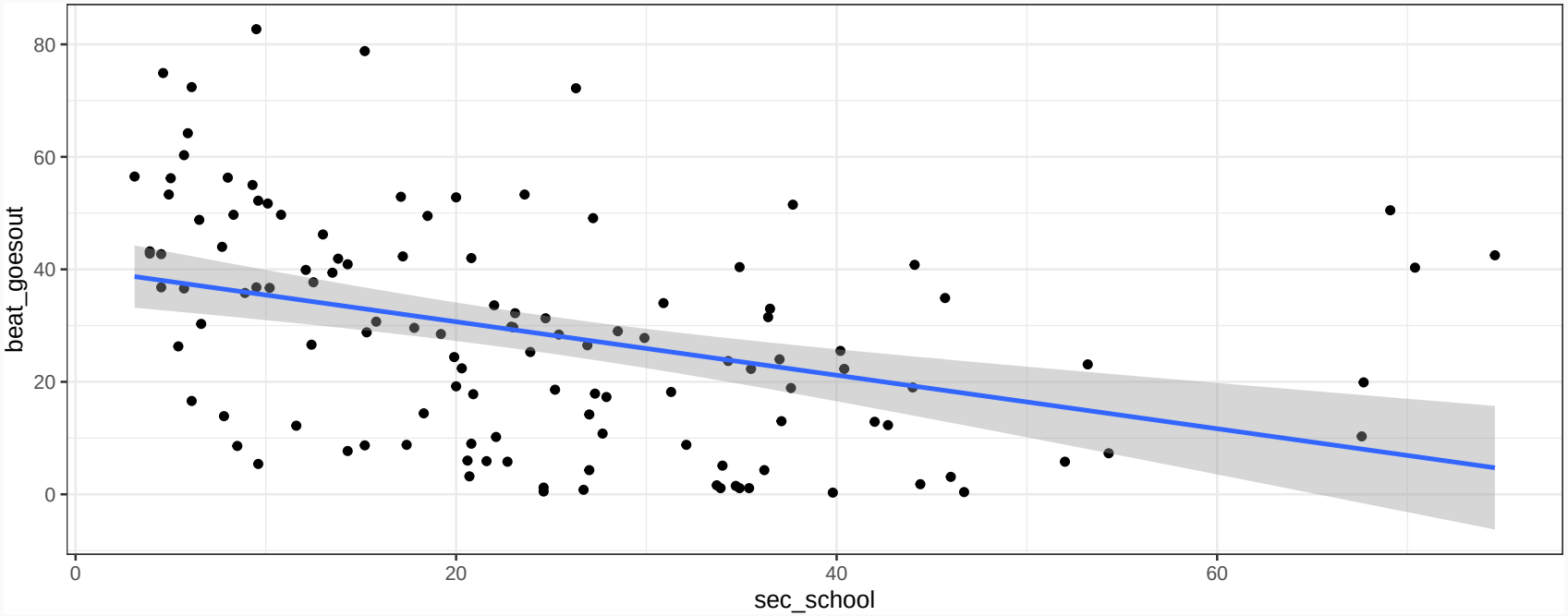


Describing linear associations: correlation

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

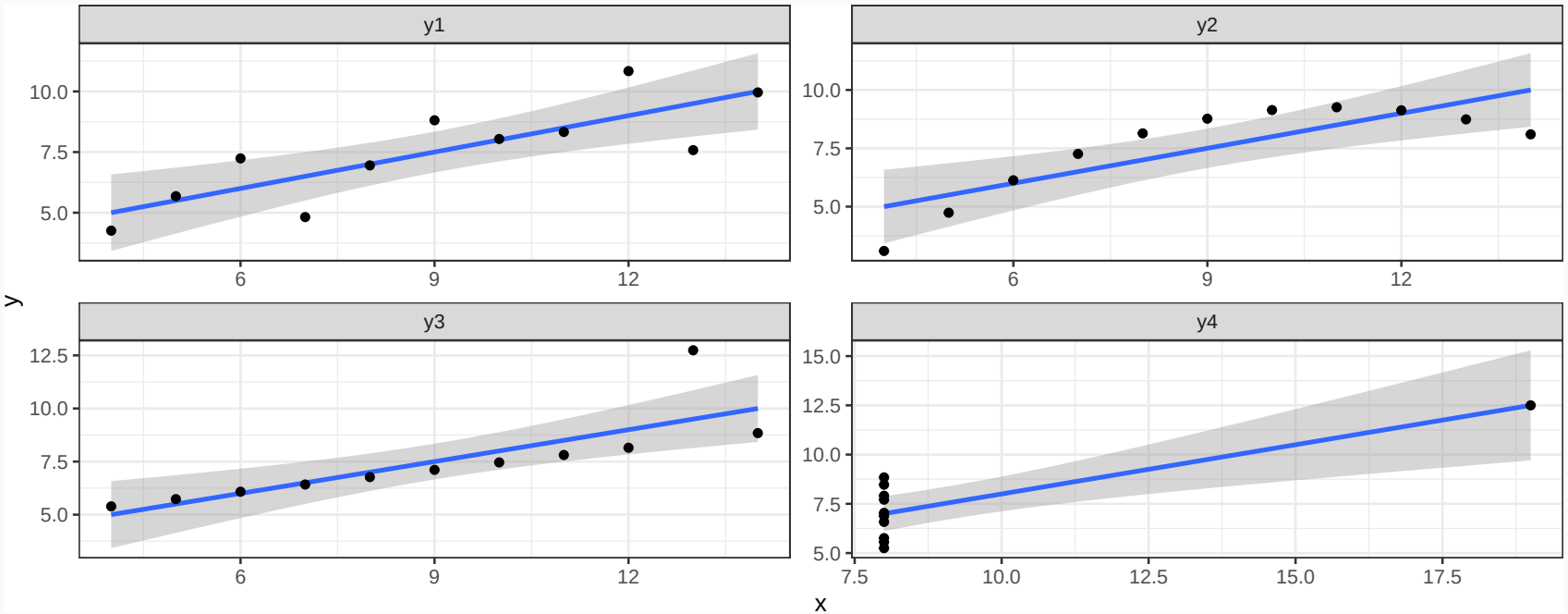
```
## [1] -0.3802336
```

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



Limits of correlation coefficients and importance of visualization

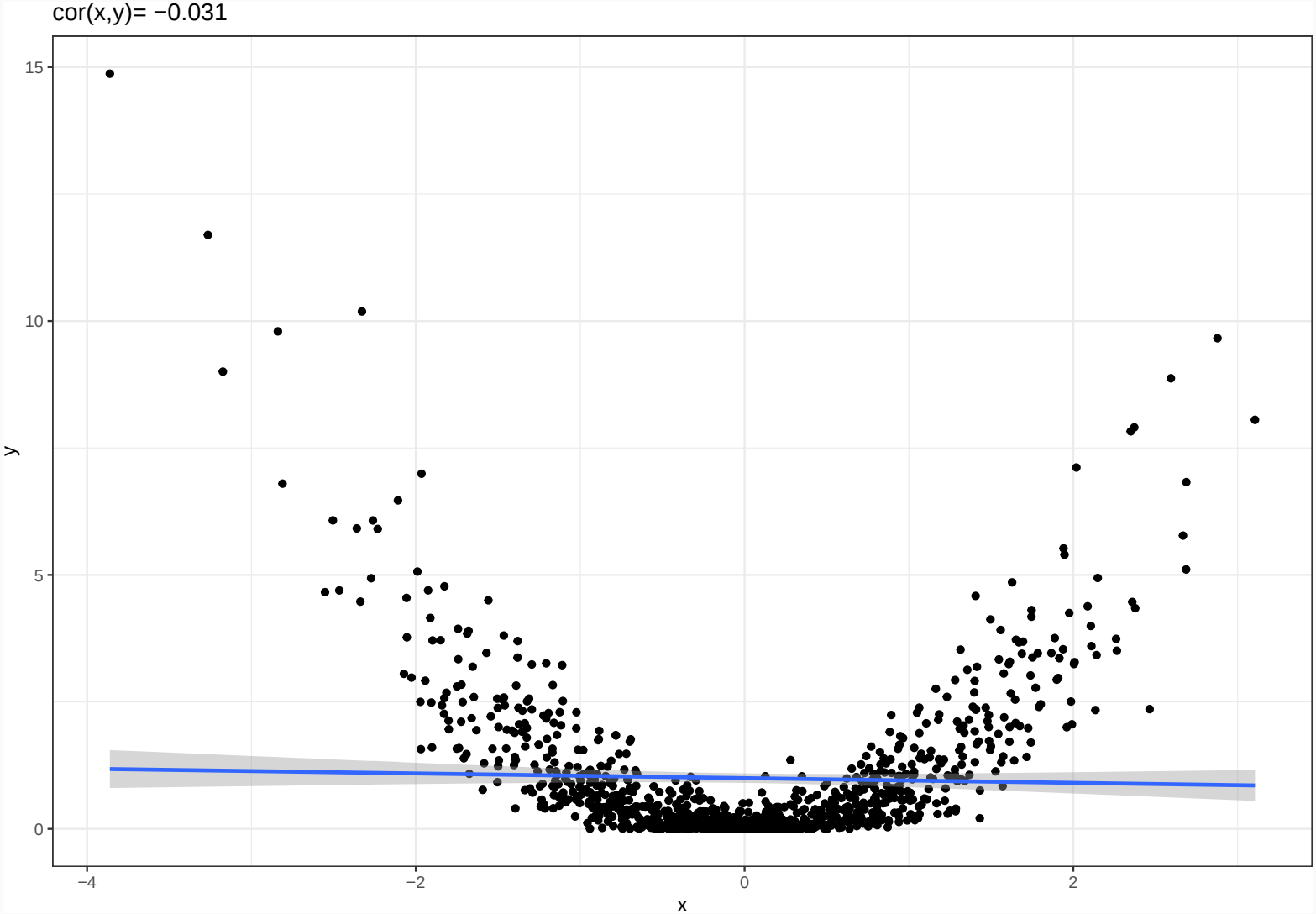
var	cor
y1	0.8164
y2	0.8162
y3	0.8163
y4	0.8165



Correlations and linear relationships

- A correlation coefficient ranges between [-1,1]
- A correlation coefficient of 1 or -1 indicates a perfect linear association: x=y
- A positive correlation coefficient indicates a positive slope
- A negative correlation coefficient indicates a negative slope
- A weak correlation does not imply that there is no relationship

Limits of linear relationships (continued)



The linear regression model

We can describe the relationship between a predictor variable X and an outcome variable Y with the line:

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

Where β_0 is the y-intercept of the line, β_1 is the slope of the line, and ε is the error between the fitted line and the coordinates (X, Y)

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.

The linear regression model and prediction

Remember, that we put a *hat* on variables to indicate that they are estimated from the data, or predicted.

In other words, we try to learn about the *regression coefficients* β_1 and β_0 by estimating $\hat{\beta}_1$ and $\hat{\beta}_0$.

A regression line predicts values Y, $\hat{Y}$ with the equation:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

and the residual, or prediction error is the difference between the observed and predicted values of Y

$$\varepsilon = Y - \hat{Y}$$

Understanding the regression line

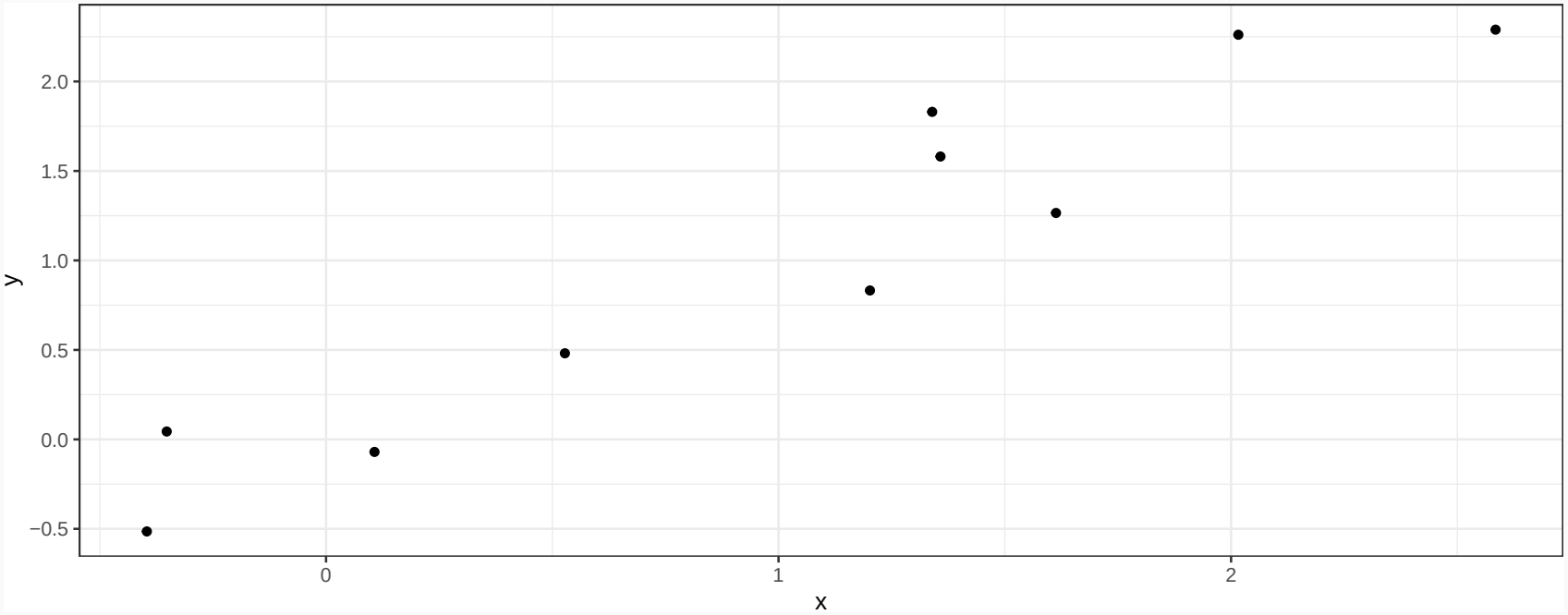
```
## # A tibble: 10 x 2
##         x       y
##     <dbl>   <dbl>
##  1  0.528  0.481
##  2 -0.396 -0.514
##  3  2.58   2.29
##  4  1.61   1.27
##  5  1.36   1.58
##  6  2.02   2.26
##  7  1.20   0.832
##  8  0.107 -0.0700
##  9 -0.352  0.0441
## 10  1.34   1.83
```

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$

- Estimate $\hat{Y}$. Recall that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
- Estimate ε. Recall that $\varepsilon = Y - \hat{Y}$

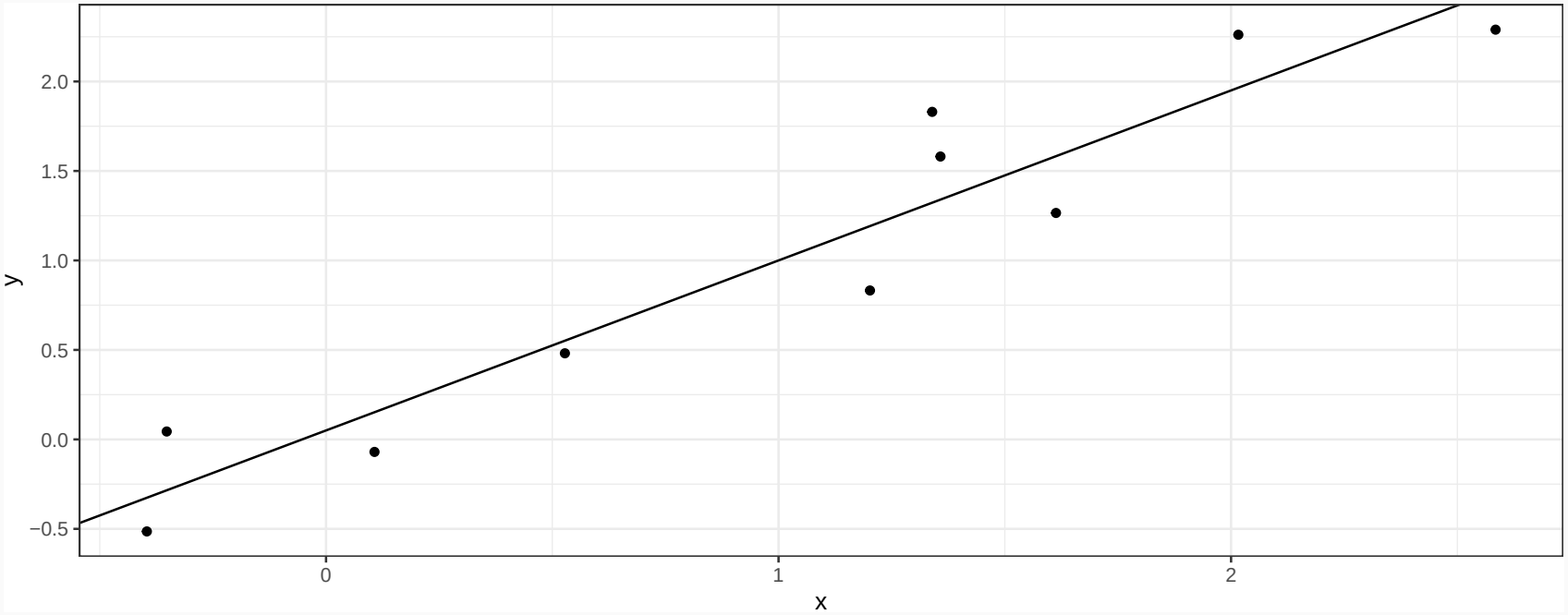
Understanding the regression line

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$



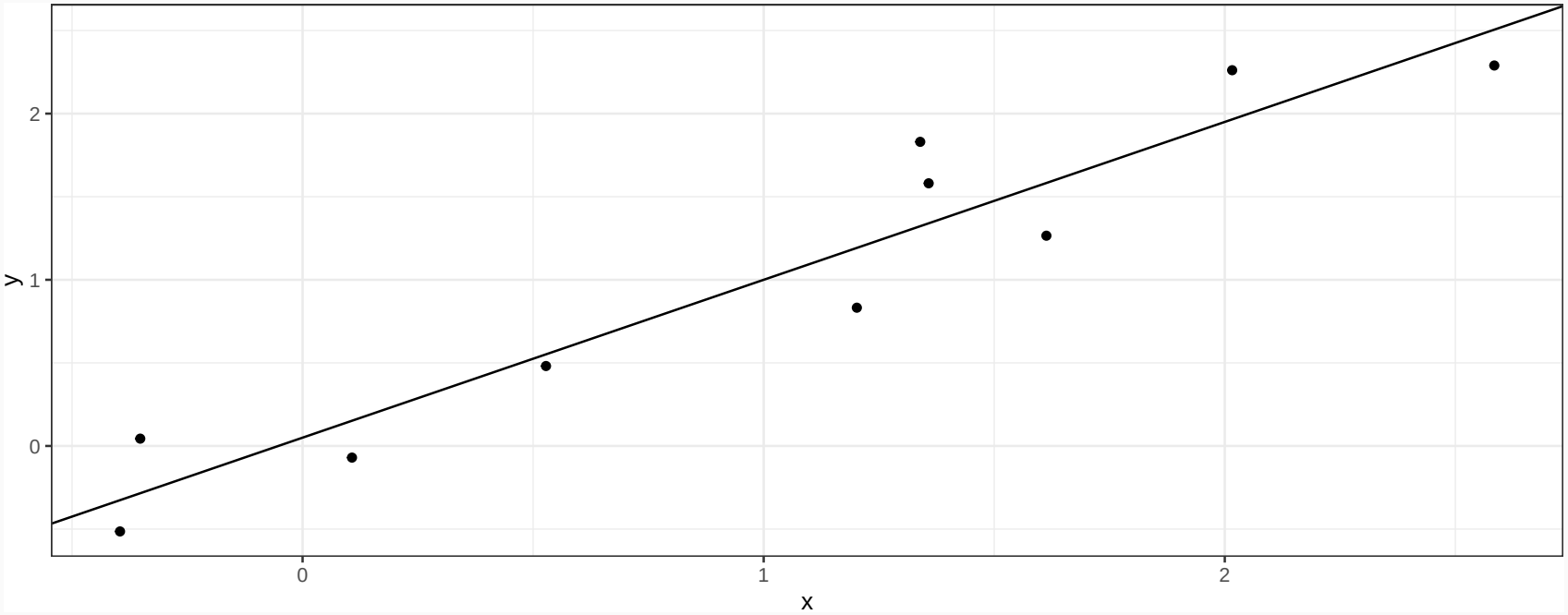
Understanding the regression line: adding the fit

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$



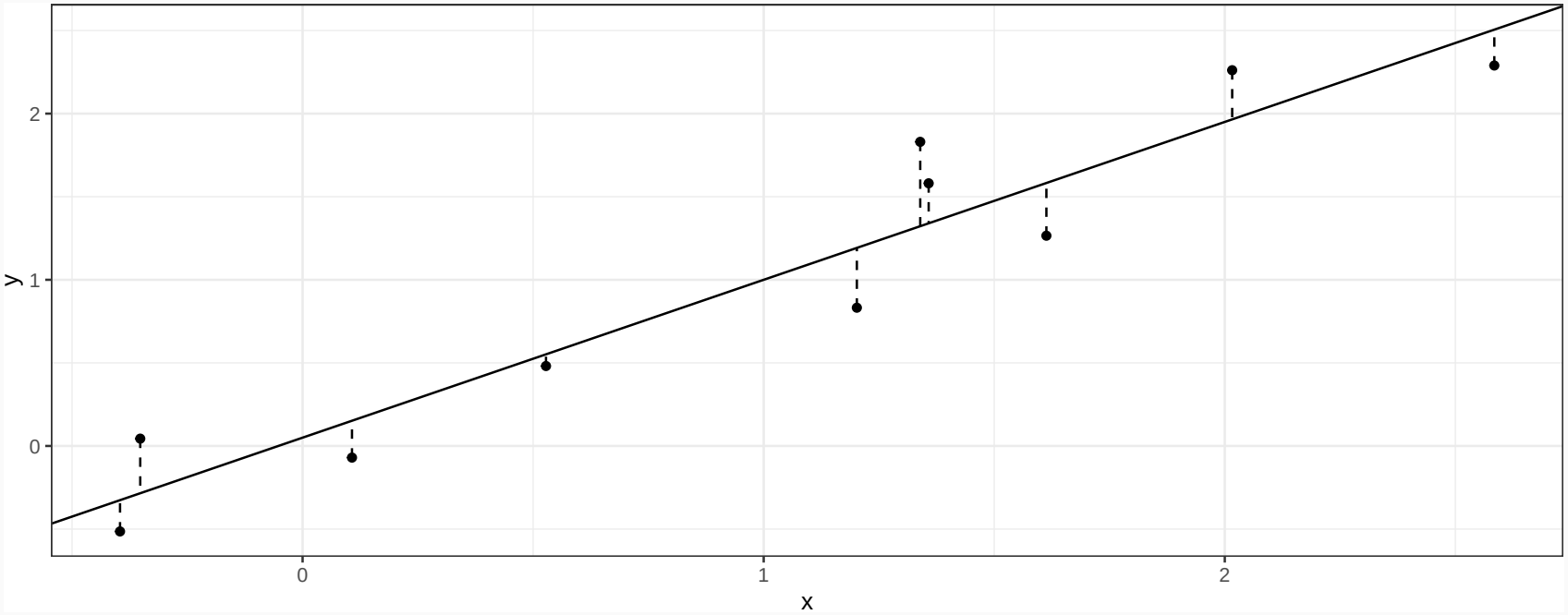
Understanding the regression line: adding $\hat{y}$

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$



Understanding the regression line: adding ε

$\hat{\beta}_0 = 0.05, \hat{\beta}_1 = 0.95$



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 $\hat{\beta}_0$ and $\hat{\beta}_1$ that results in the smallest possible value for:

$$\text{SSR} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X)^2$$

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

$$\hat{\beta}_1 = (X^{\mathrm{T}} X)^{-1} X^{\mathrm{T}} Y$$

See Imai for a more details on the math behind OLS

Estimating a regression model in R, the basics

```r
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
```

Estimating a linear regression model in R, IPV data

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

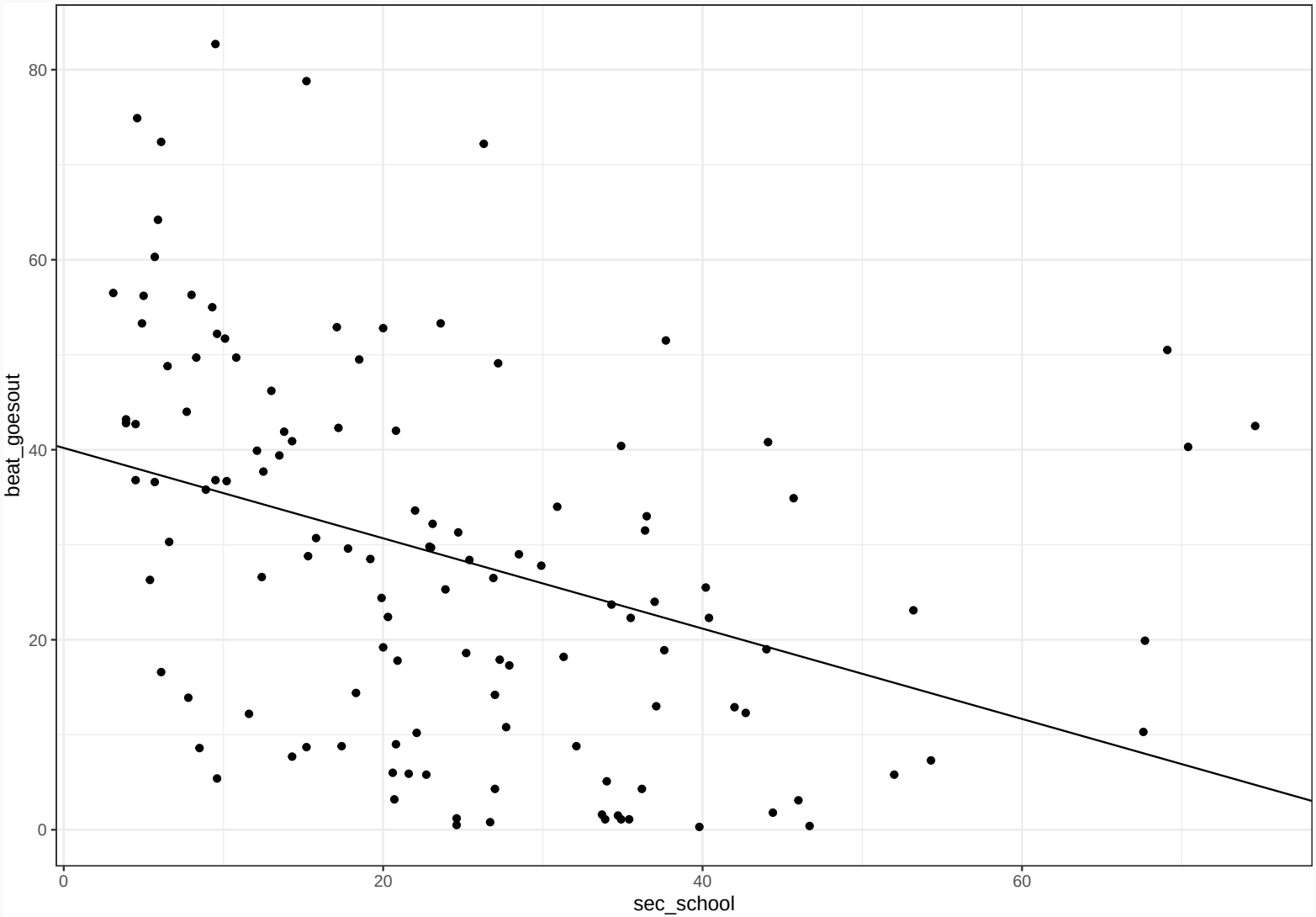
```
ipv_model <- lm(beat_goesout ~ sec_school, data = ipv)

coef(ipv_model)
```

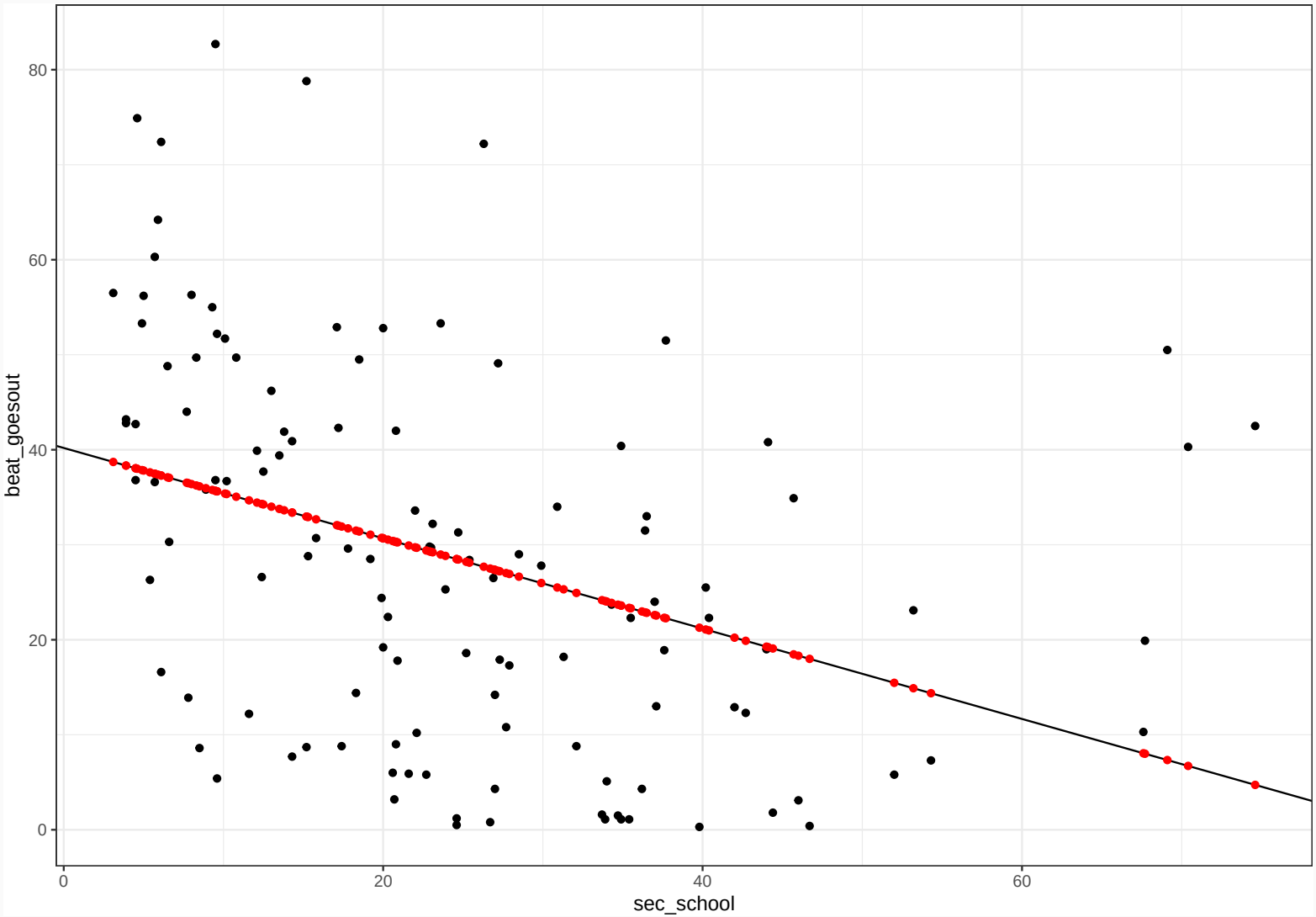
```
## (Intercept)  sec_school
##  40.1876597  -0.4753799
```

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

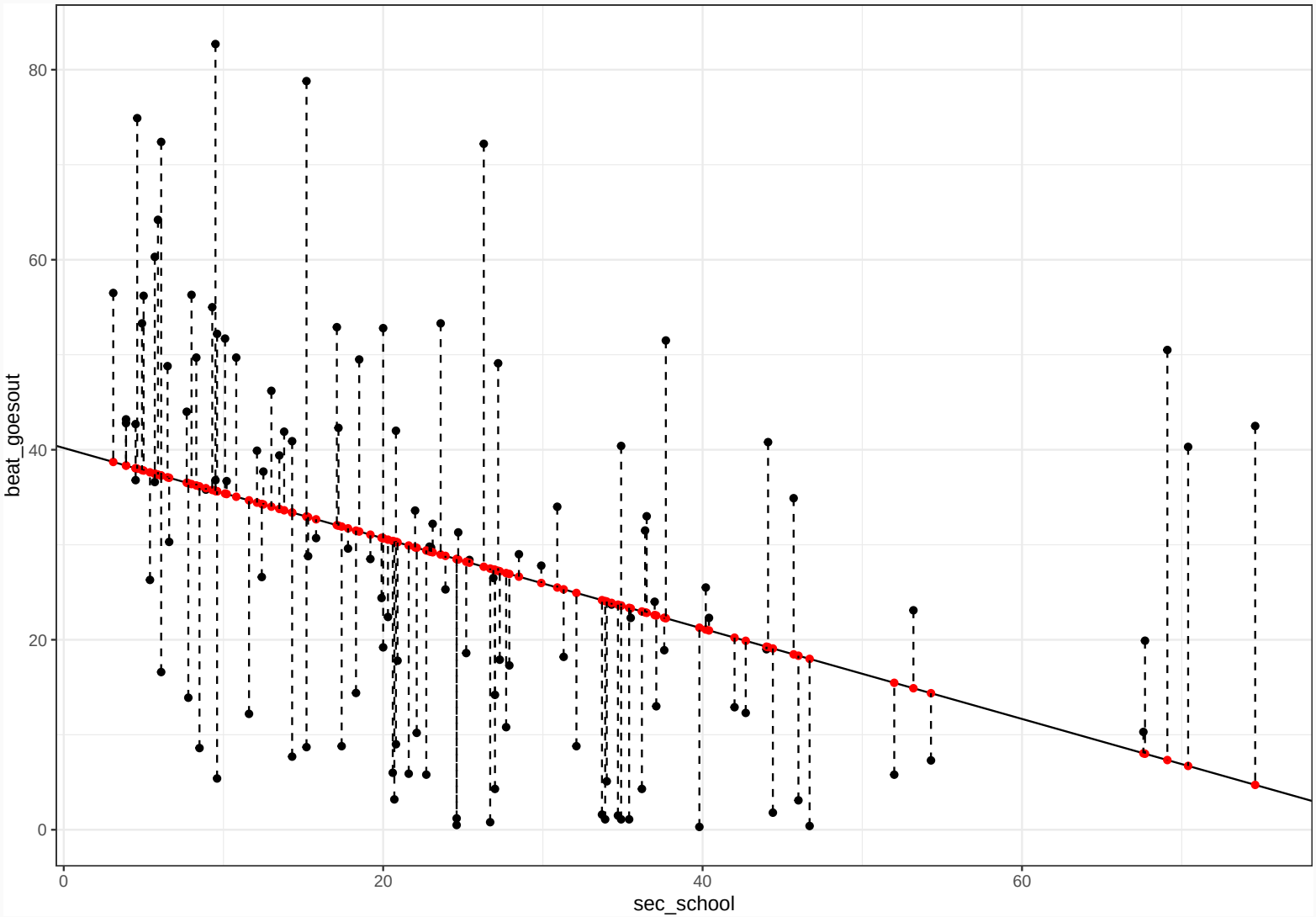
Visualize the model



Visualize the model



Visualize the model



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.

Consistent with our expectations, there is a negative linear relationship between secondary schooling and women's attitudes about intimate partner violence.

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:

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

To be more compact, I often write this in matrix notation as

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

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: one continuous, one categorical

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

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

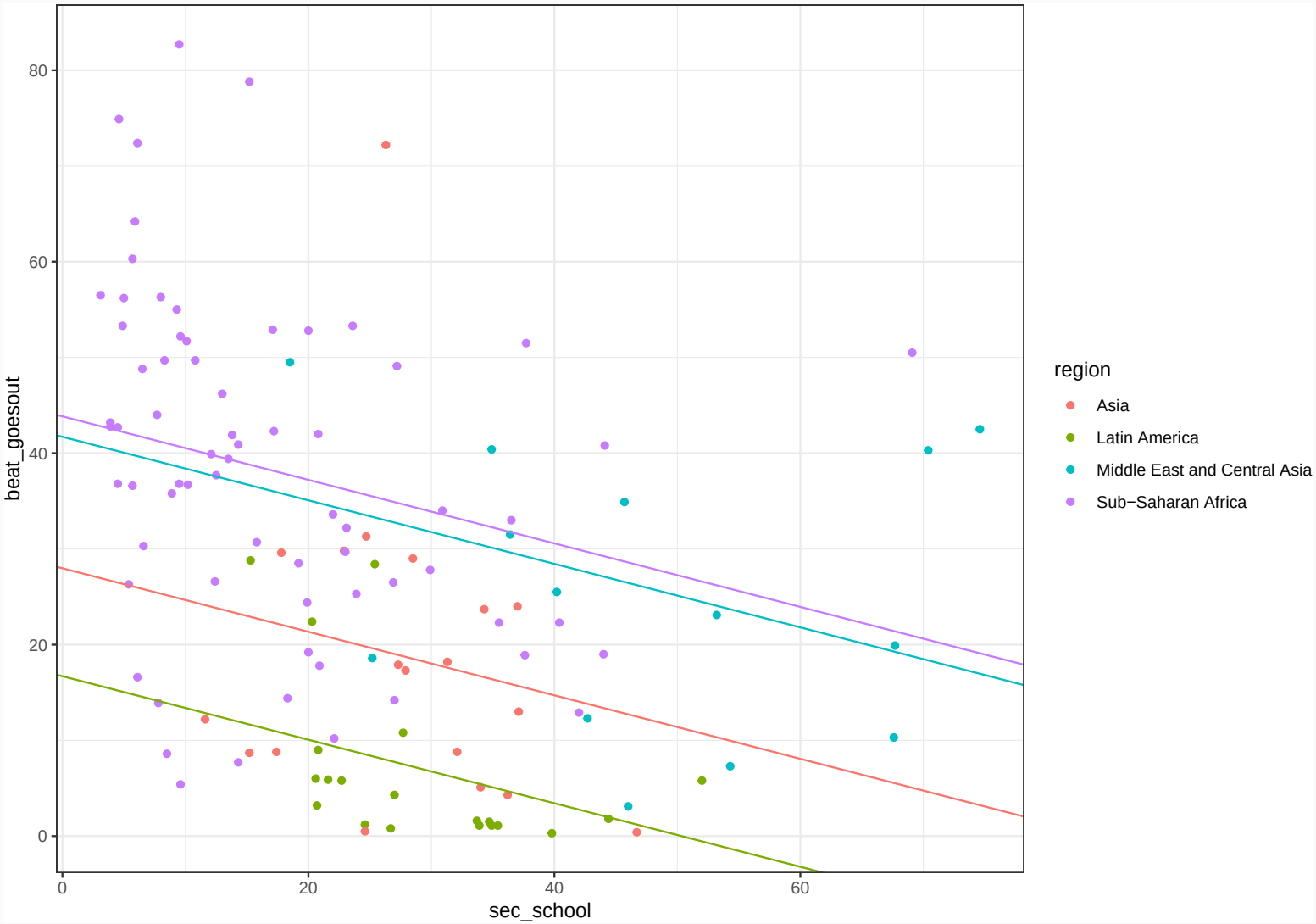
Let's add a predictor for region. Remember from prior examples that we saw clear patterns within regions.

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

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)

coef(ipv_model3)
```

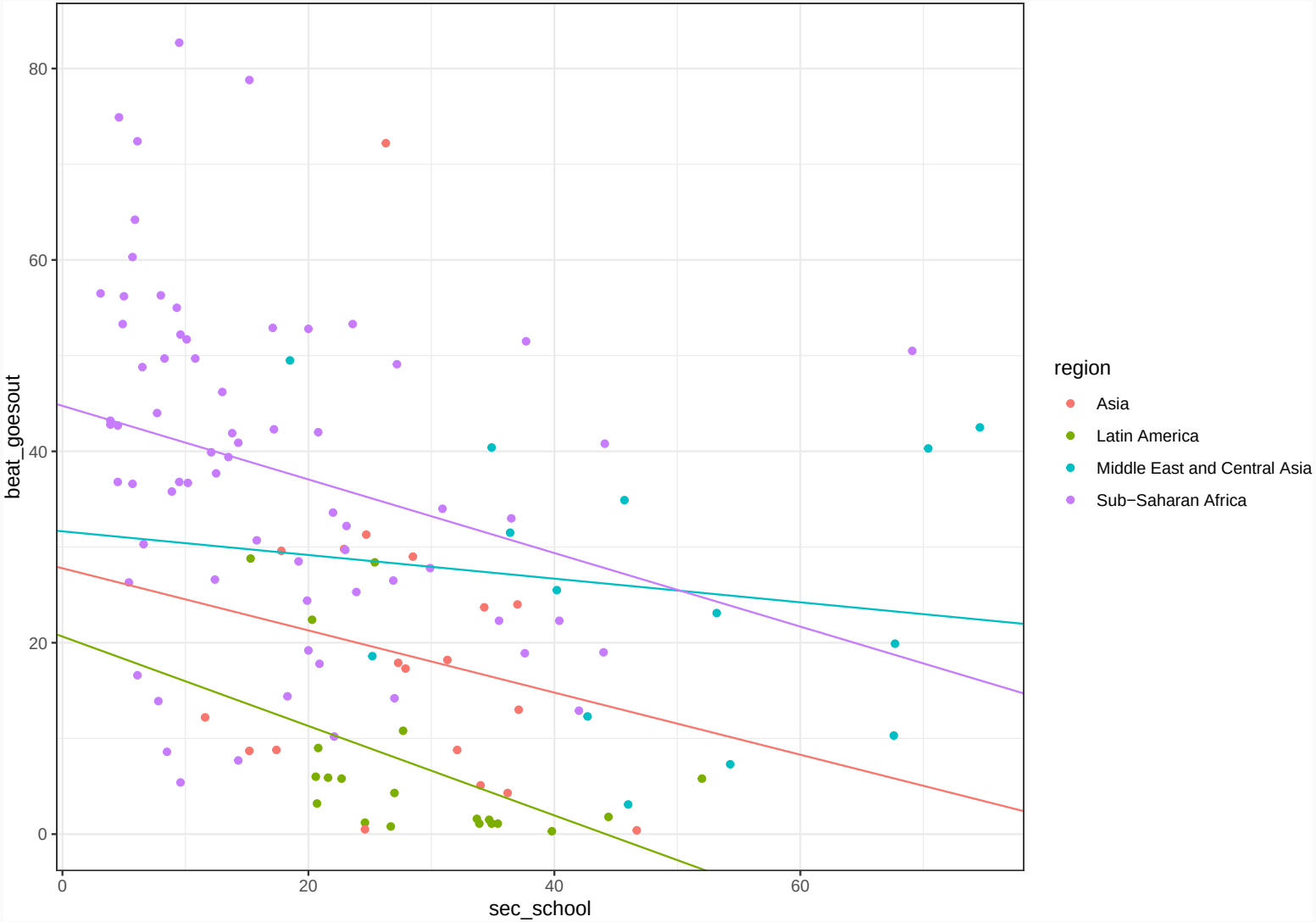
```
##                                     (Intercept)
##                                     27.78048328
##                                      sec_school
##                                     -0.32469353
##                                regionLatin America
##                                     -7.13303634
##             regionMiddle East and Central Asia
##                                      3.85875152
##                           regionSub-Saharan Africa
##                                     16.97257959
##                    sec_school:regionLatin America
##                                     -0.14258393
## sec_school:regionMiddle East and Central Asia
##                                      0.20106010
##               sec_school:regionSub-Saharan Africa
##                                     -0.05977263
```

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, and hence takes the unmodified version of the intercept and slope.

Visualizing interactions



Fitting transformed predictors

What if we think that the relationship between secondary schooling and tolerance for IPV is non-linear? Maybe there are large decreases in tolerance for increases in sec_school when it is near zero, but diminishing decreases later?

```
ipv_model4 <- lm(beat_goesout ~ I(sec_school^2) + region + region * s
    data = ipv)
```

Use I() when doing math to a predictor in a model. No need to transform beforehand.

Common transformations also include (I() not required for these):

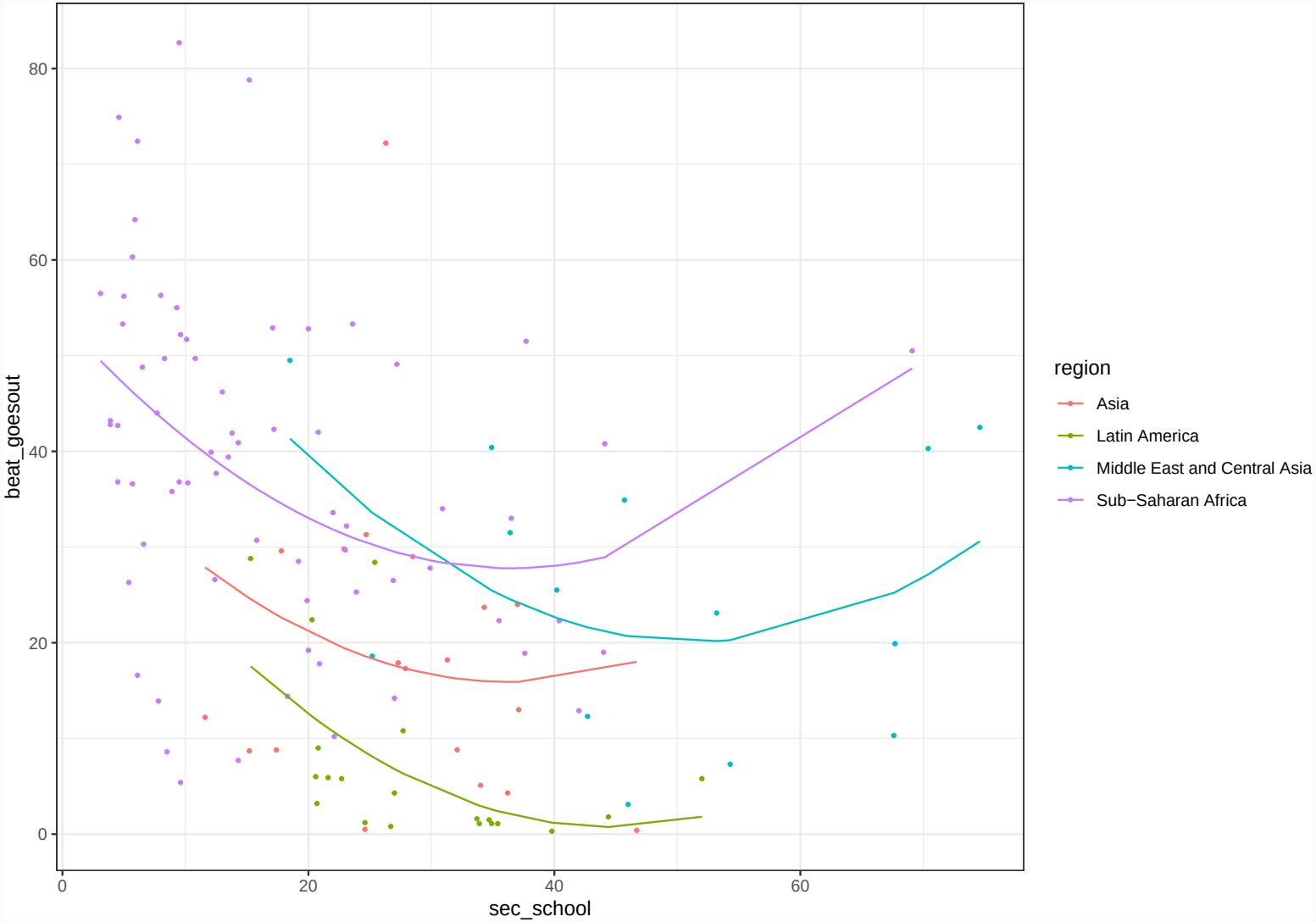
- log()
- sqrt()
- scale()

Look at the model

```
coef(ipv_model4)
```

```
##                                     (Intercept)
##                                    41.726281036
##                                  I(sec_school^2)
##                                     0.019553566
##                               regionLatin America
##                                    -2.083742069
##          regionMiddle East and Central Asia
##                                    30.130931486
##                        regionSub-Saharan Africa
##                                    11.967087542
##                                      sec_school
##                                    -1.421176316
##                    regionLatin America:sec_school
##                                    -0.323073123
## regionMiddle East and Central Asia:sec_school
```

Visualize the model



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.

Lab

- Causal inference using linear regression models (RCT, regression discontinuity examples)

Homework

- Question 4.5.2. Due by 10AM on Wednesday 10/23 (aka the start of lecture)