Prediction, 1

Frank Edwards

11/2/21

Prediction

Why predict?

- To learn about what is likely to happen in the future (weather, elections, economic changes)

Why predict?

- To learn about what is likely to happen in the future (weather, elections, economic changes)
- To validate theories or arguments:
    - Valid causal inference requires successful prediction of counterfactual claims

Why predict?

- To learn about what is likely to happen in the future (weather, elections, economic changes)
- To validate theories or arguments:
    - Valid causal inference requires successful prediction of counterfactual claims
    - e.g. if X were different, what value of Y would we observe?

Load and process polls data

```
polls<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/polls2016.csv")

results<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/1976-2016-president.csv"
```

How can we join these two data frames?

- Harmonize the data (consistent column names, variable labels, units of analysis)

How can we join these two data frames?

- Harmonize the data (consistent column names, variable labels, units of analysis)

```
polls<-polls %>%
  filter(population == "Likely Voters") %>%
  select(state, electoral_votes, Clinton, Trump) %>%
  pivot_longer(cols = Clinton:Trump,
               names_to = "candidate",
               values_to = "poll_result")

results<-results %>%
  filter(year==2016,
         office == "US President",
         party == "democrat" | party == "republican") %>%
  mutate(candidate = case_when(
    candidate == "Clinton, Hillary" ~ "Clinton",
    candidate == "Trump, Donald J." ~ "Trump"
  )) %>%
  select(state_po, candidate, candidatevotes, totalvotes) %>%
  rename(state = state_po)
```

How can we join these two data frames?

- Join the data frames
- create needed variables for analysis

```
polls_results<-polls %>%
  left_join(results) %>%
  mutate(pct_vote = candidatevotes / totalvotes * 100)
```

Calculate prediction error

Error is a general term for how wrong our guess is. We can generally calculate error by subtracting the observation from our prediction.

Calculate prediction error

Error is a general term for how wrong our guess is. We can generally calculate error by subtracting the observation from our prediction.

`prediction error` = `predicted value` - `observed value`

Calculate prediction error

Error is a general term for how wrong our guess is. We can generally calculate error by subtracting the observation from our prediction.

$$\texttt{prediction error} = \texttt{predicted value} - \texttt{observed value}$$

```r
polls_results<-polls_results %>%
  mutate(error = poll_result - pct_vote)

head(polls_results)
```

```
## # A tibble: 6 x 8
##   state electoral_votes candidate poll_result candidatevotes totalvotes pct_vote
##   <chr>           <int> <chr>           <int>          <int>      <int>    <dbl>
## 1 TX                 38 Clinton            38        3877868    8969226     43.2
## 2 TX                 38 Trump              41        4685047    8969226     52.2
## 3 WI                 10 Clinton            48        1382536    2976150     46.5
## 4 WI                 10 Trump              44        1405284    2976150     47.2
## 5 VA                 13 Clinton            54        1981473    3982752     49.8
## 6 VA                 13 Trump              41        1769443    3982752     44.4
## # ... with 1 more variable: error <dbl>
```

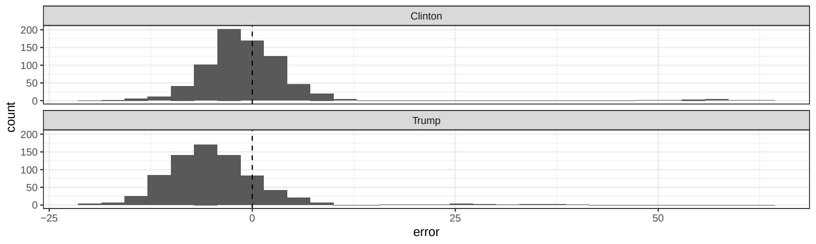
Evaluate the errors

```r
polls_results %>%
  group_by(candidate) %>%
  summarise(mean_error = mean(error))
```

```
## # A tibble: 2 x 2
##   candidate mean_error
##   <chr>          <dbl>
## 1 Clinton       -0.381
## 2 Trump         -4.64
```

Evaluate the errors

```
ggplot(polls_results,
       aes(x = error)) +
  geom_histogram() +
  geom_vline(aes(xintercept = 0), lty=2) +
  facet_wrap(~candidate, ncol = 1)
```



Root Mean Square Error

RMSE provides a measure of absolute error, where positive and negative errors don’t negate each other

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - y)^2}{n}}$$

```
polls_results %>%
  group_by(candidate) %>%
  summarise(rmse = sqrt(mean(error^2)))
```

```
## # A tibble: 2 x 2
##   candidate  rmse
##   <chr>     <dbl>
## 1 Clinton    8.21
## 2 Trump      8.28
```

Conclusions on errors

1. Polls had similar magnitude of error for both candidates (RMSE)
2. Poll errors were consistently negative for Trump, were zero on average for Clinton.

How many polls called it right?

1. Make an average prediction for each state across polls
2. Whichever candidate has the highest average polling number is predicted the winner

Making a prediction based on the polls

```r
poll_winner<-polls_results %>%
  group_by(state, candidate) %>%
  summarise(poll_mean = mean(poll_result)) %>%
  filter(poll_mean == max(poll_mean))

# group_by with filter will perform the filter operation
# over each unit in the group (states)

table(poll_winner$candidate)
```

```
##
## Clinton   Trump
##      26      24
```

What percent of electoral college votes does our prediction yield for Clinton

```
poll_winner %>%
  left_join(polls %>%
            select(state, electoral_votes) %>%
            distinct()) %>%
  group_by(candidate) %>%
  summarise(electoral_pct_polls = sum(electoral_votes) / 538 * 100)
```

```
## # A tibble: 2 x 2
##   candidate electoral_pct_polls
##   <chr>                   <dbl>
## 1 Clinton                  62.8
## 2 Trump                    36.6
```

```
## actual result for Clinton
227/538 * 100
```

```
## [1] 42.19331
```

Classification: potential outcomes for binary predictions

Bold cells are correct classifications.

	Positive, obs.	Negative, obs.
Positive, pred.	**True positive**	False positive
Negative, pred.	False negative	**True negative**

Check our performance

- First, format the data to join the election winner onto our poll result predictions

```
poll_winner<-poll_winner %>%
  select(state, candidate) %>%
  rename(poll_winner = candidate)

election_winner<-results %>%
  group_by(state) %>%
  filter(candidatevotes == max(candidatevotes)) %>%
  rename(election_winner = candidate) %>%
  select(state, election_winner)

head(election_winner)
```

```
## # A tibble: 6 x 2
## # Groups:   state [6]
##   state election_winner
##   <chr> <chr>
## 1 AL    Trump
## 2 AK    Trump
## 3 AZ    Trump
## 4 AR    Trump
## 5 CA    Clinton
## 6 CO    Clinton
```

```
head(poll_winner)
```

Join the data frames

```
poll_performance<-poll_winner %>%
  left_join(election_winner)

head(poll_performance)
```

```
## # A tibble: 6 x 3
## # Groups:   state [6]
##   state poll_winner election_winner
##   <chr> <chr>       <chr>
## 1 AK    Trump       Trump
## 2 AL    Trump       Trump
## 3 AR    Trump       Trump
## 4 AZ    Trump       Trump
## 5 CA    Clinton     Clinton
## 6 CO    Clinton     Clinton
```

Then make correct classification a binary outcome

```
polls_performance<-poll_performance%>%
  mutate(poll_correct = poll_winner == election_winner)
```

How often were the polls right?

```
## calculate proportion of accurate classifications
## i.e. clinton_wins_pred == clinton_wins_vote

mean(polls_performance$poll_correct)
```

```
## [1] 0.88
```

Which ones did they get wrong?

```r
## Get misclassifications
polls_performance %>%
  filter(!poll_correct)
```

```
## # A tibble: 6 x 4
## # Groups:   state [6]
##   state poll_winner election_winner poll_correct
##   <chr> <chr>       <chr>           <lgl>
## 1 FL    Clinton     Trump           FALSE
## 2 MI    Clinton     Trump           FALSE
## 3 NC    Clinton     Trump           FALSE
## 4 OH    Clinton     Trump           FALSE
## 5 PA    Clinton     Trump           FALSE
## 6 WI    Clinton     Trump           FALSE
```

Linear regression

Linear regression: IPV data

```
ipv<-read.csv("https://raw.githubusercontent.com/f-edwards/intro_stats/master/data/dhs_ipv.csv")

head(ipv)
```

```
##   X beat_burnfood beat_goesout sec_school no_media    country year
## 1 1           4.4         18.6       25.2      1.5    Albania 2008
## 2 4           4.9         19.9       67.7      8.7    Armenia 2000
## 3 5           2.1         10.3       67.6      2.2    Armenia 2005
## 4 6           0.3          3.1       46.0      6.4    Armenia 2010
## 5 7          12.1         42.5       74.6      7.4 Azerbaijan 2006
## 6 8            NA           NA       24.0     41.9 Bangladesh 2004
##                          region
## 1 Middle East and Central Asia
## 2 Middle East and Central Asia
## 3 Middle East and Central Asia
## 4 Middle East and Central Asia
## 5 Middle East and Central Asia
## 6                         Asia
```

Research question

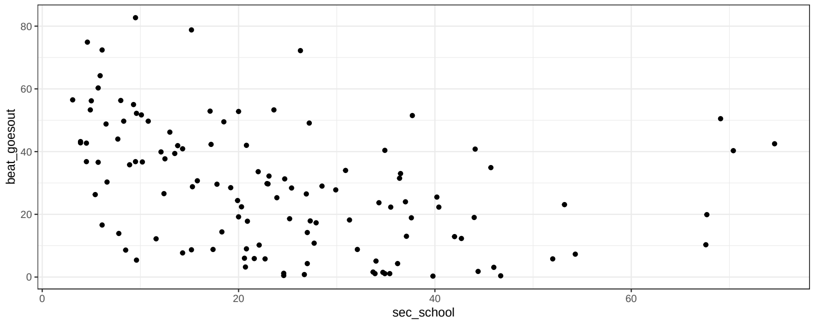
- Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?

Research question

- Are secondary school completion rates for women associated with lower levels of acceptance of intimate partner violence?
- *Hypothesis:* Feminist theory suggests a negative association between schooling and tolerance for intimate partner violence. In places where women have more social and economic power, tolerance for intimate partner violence should be lower.

Visualizing associations: scatterplots

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

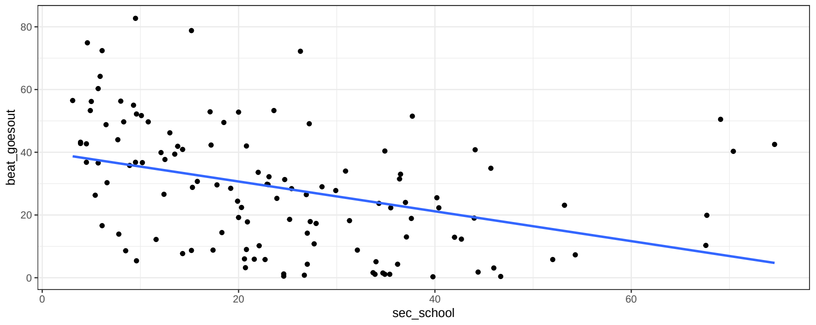


Describing linear associations: correlation

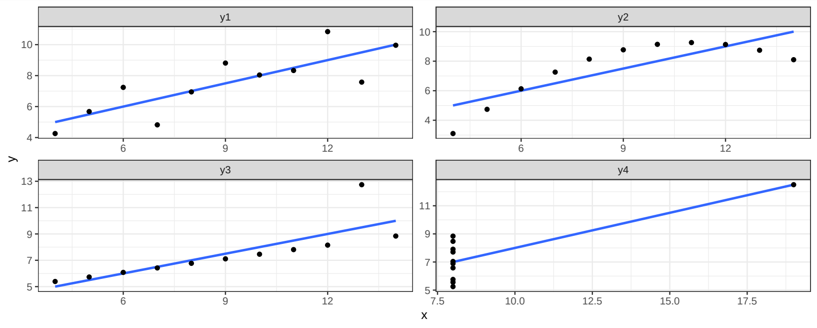
```
cor(ipv$sec_school, ipv$beat_goesout, use = "complete")
```

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

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

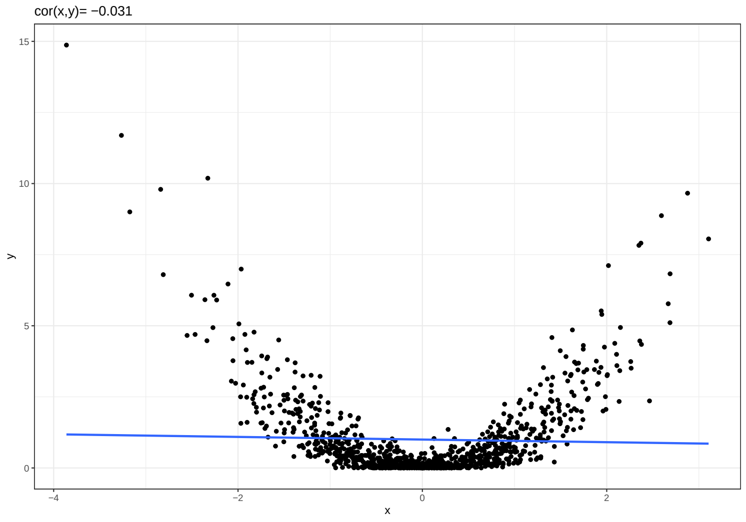


Limits of correlation coefficients and importance of visualization




```
## # A tibble: 4 x 2
##   var     cor
##   <chr> <dbl>
## 1 y1    0.816
## 2 y2    0.816
## 3 y3    0.816
## 4 y4    0.817
```

Limits of linear relationships (continued)



- A correlation coefficient ranges between [-1,1]

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 (if x and y are SD scaled)

- A correlation coefficient ranges between [-1,1]
- A correlation coefficient of 1 or -1 indicates a perfect linear association: x=y (if x and y are SD scaled)
- A positive correlation coefficient indicates a positive slope

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 (if x and y are SD scaled)
- A positive correlation coefficient indicates a positive slope
- A negative correlation coefficient indicates a negative slope

- A correlation coefficient ranges between [-1,1]
- A correlation coefficient of 1 or -1 indicates a perfect linear association: x=y (if x and y are SD scaled)
- 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

Lines

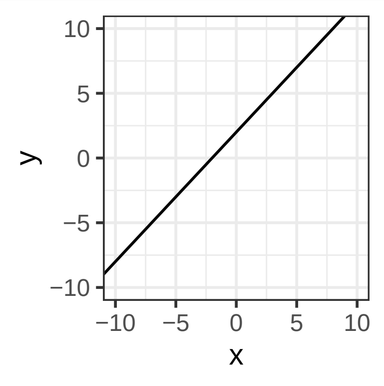
We can define a line as:

$$y = mx + b$$

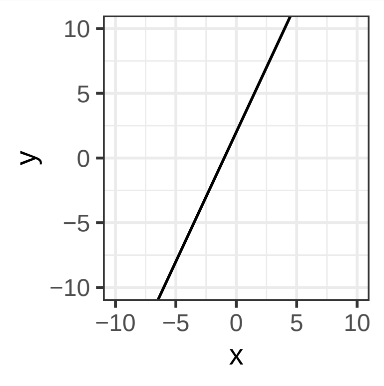
Where m is the slope and b is the y-intercept.

What will the line $y = x + 2$ look like?

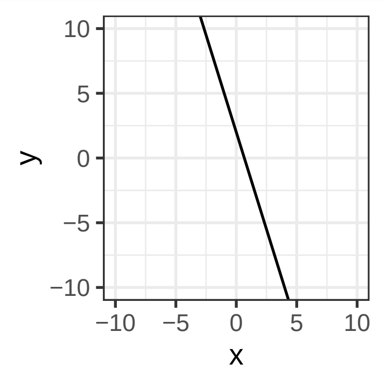
What will the line $y = x + 2$ look like?



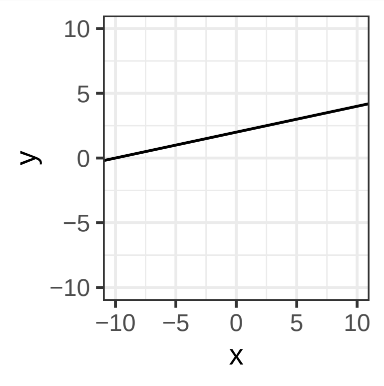
What will the line $y = 2x + 2$ look like?



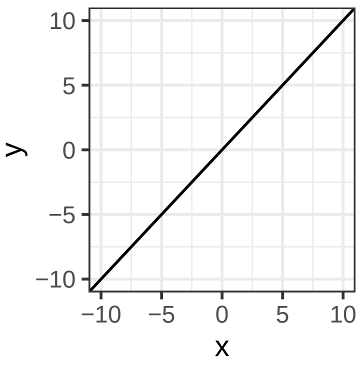
What will the line $y = -2x + 2$ look like?



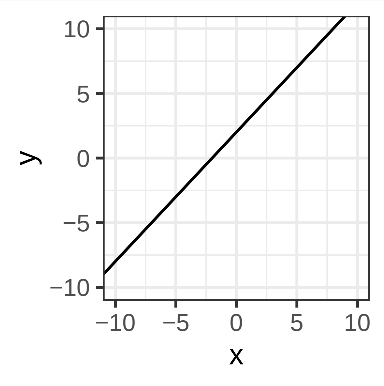
What will the line $y = 0.2x + 2$ look like?



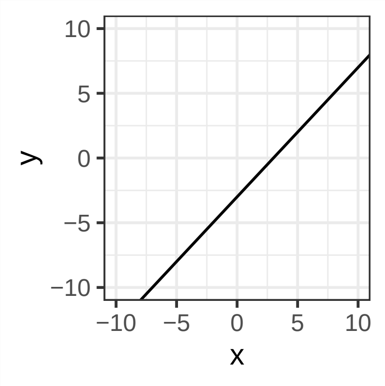
What will the line $y = x + 0$ look like?



What will the line $y = x + 2$ look like?



What will the line $y = x - 3$ look like?



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

$$y = mx + b$$

What does m describe?

The linear regression model

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

$$y = mx + b$$

What does m describe?

- The increase in y for a one-unit increase in x

What does b describe

The linear regression model

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

$$y = mx + b$$

What does m describe?

- The increase in y for a one-unit increase in x

What does b describe

- The location of y when x = 0

The linear regression model: expected value

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

$$E[Y] = \beta_0 + \beta_1 X$$

The linear regression model: expected value

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

$$E[Y] = \beta_0 + \beta_1 X$$

What does β_0 describe?

The linear regression model: expected value

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

$$E[Y] = \beta_0 + \beta_1 X$$

What does β_0 describe?

What does β_1 describe?

The error term in linear regression

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

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

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

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.

The linear regresion model as a prediction engine

The line $y_i = \beta_0 + \beta_1 x_i$ provides a prediction for the value y_i based on the value of x_i.

The linear regresion model as a prediction engine

The line $y_i = \beta_0 + \beta_1 x_i$ provides a prediction for the value y_i based on the value of x_i.

- If $\beta_0 = 2$ and $\beta_1 = 1.5$, what is the expected value of y when $x = 4$?

The linear regresion model as a prediction engine

The line $y_i = \beta_0 + \beta_1 x_i$ provides a prediction for the value y_i based on the value of x_i.

- If $\beta_0 = 2$ and $\beta_1 = 1.5$, what is the expected value of y when $x = 4$?
- When $x = 2$?

The linear regression model and prediction

We put a $\hat{hat}$ on variables to indicate that they are estimated from the data, or predicted.

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

$$\hat{Y} = \beta_0 + \beta_1 X$$

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

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

Understanding the regression line for real data

```
## # 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
```

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

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

Understanding the regression line for real data

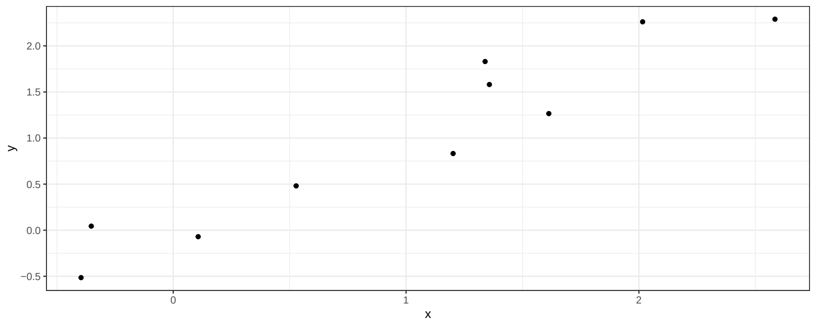
```
## # A tibble: 10 x 2
##         x       y
##     <dbl>   <dbl>
##  1  0.528  0.481
##  2 -0.396 -0.514
##  3  2.58   2.29
##  4  1.61   1.27
##  5  1.36   1.58
##  6  2.02   2.26
##  7  1.20   0.832
##  8  0.107 -0.0700
##  9 -0.352  0.0441
## 10  1.34   1.83
```

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

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

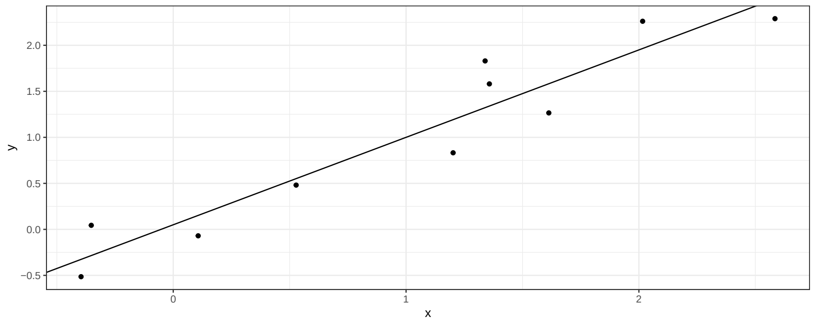
Understanding the regression line

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



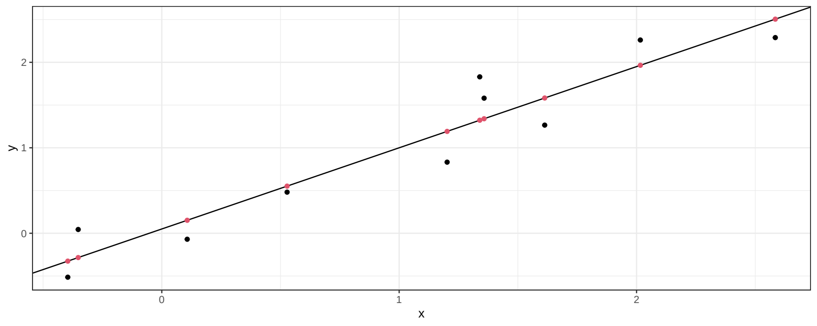
Understanding the regression line: adding the fit

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



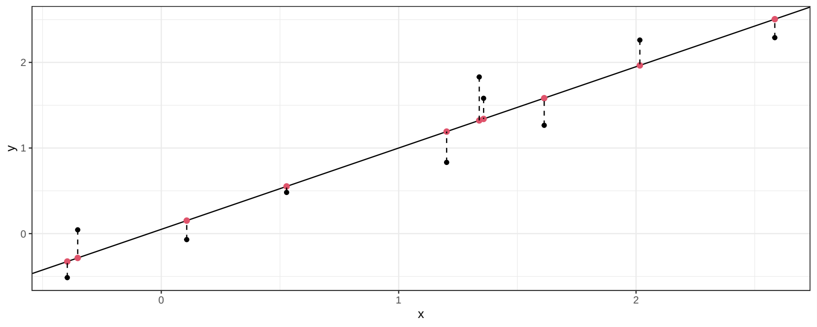
Understanding the regression line: adding $\hat{y}$

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



Understanding the regression line: adding ε

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



Ordinary least squares

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

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.

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:

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

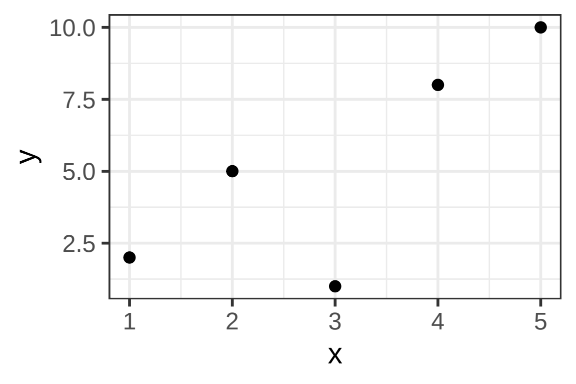
Estimating a regression model in R, the basics

```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
```

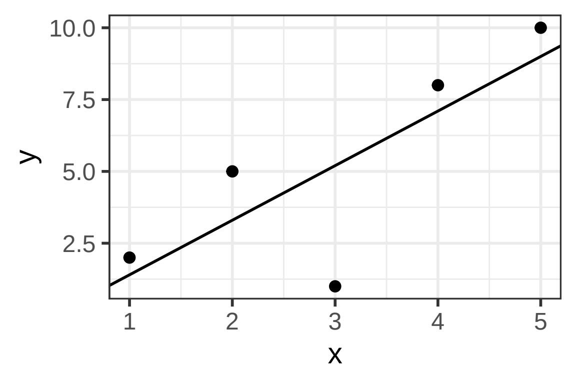
The observed data

```
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

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

```r
## 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?

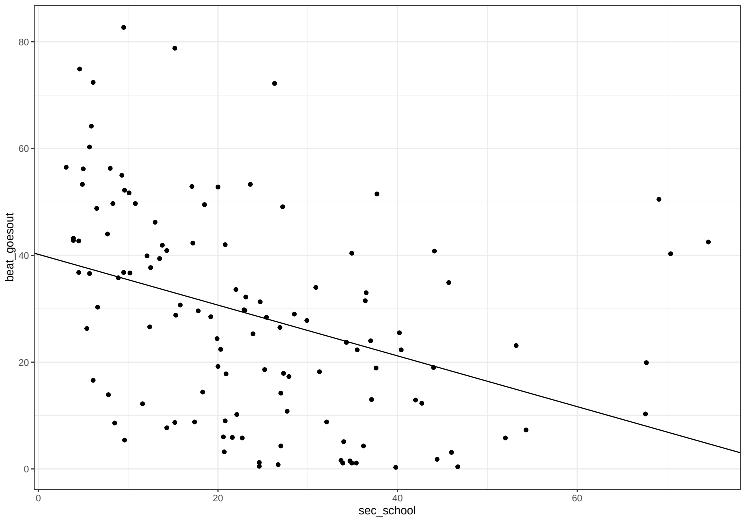
Estimating a regression model in R

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

coef(ipv_model)
```

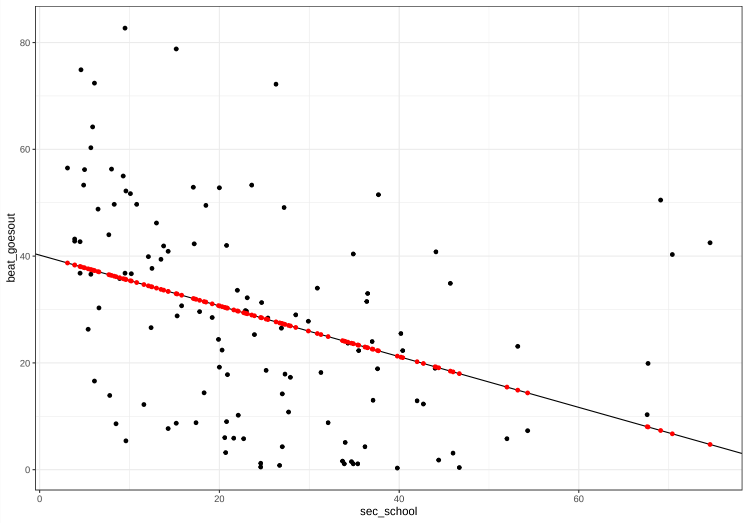
```
## (Intercept)  sec_school
##  40.1876597  -0.4753799
```

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

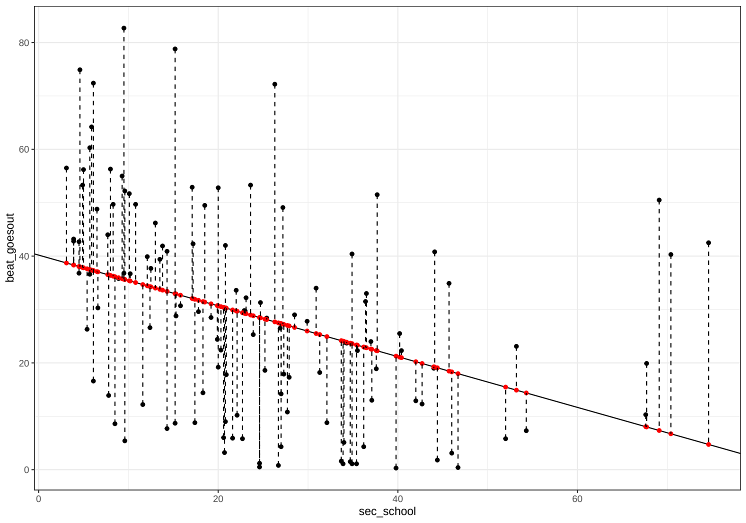
Visualize the model



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.