BAK3: Introduction to Quantitative Methods

Week 11: Bivariate Linear Regression

Leonardo Carella

The Plan for today

- Statistics:
  - Recap: Hypothesis Tests and Inferential Statistics.
  - Simple (Bivariate) **Linear Regression**.
- Coding in R:
  - Merging Dataframes.
  - Simple (Bivariate) **Linear Regression**.

```
> t.test(data$religiosity ~ data$gender)

	Welch Two Sample t-test

data:  data$religiosity by data$gender
t = -3.3561, df = 309.28, p-value = 0.0008891
alternative hypothesis: true difference in means between
group Men and group Women is not equal to 0
95 percent confidence interval:
 -0.8637381 -0.2252634
sample estimates:
  mean in group Men mean in group Women
           4.994440            5.538941
```

Is there a significant difference in religiosity (0-10 scale) between men and women?

```
> t.test(data$number_of_children ~ data$origin)
Welch Two Sample t-test

data:  data$number_of_children by data$origin
t = -0.45917, df = 129.11, p-value = 0.6469
alternative hypothesis: true difference in means between group 'Born
Abroad' and group 'Born in Austria' is not equal to 0
95 percent confidence interval:
 -0.4488909  0.2797818
sample estimates:
    mean in group 'Born Abroad' mean in group 'Born in Austria'
                       2.090909                        2.175464
```

Is there a significant difference in number of children between people born in Austria and people born abroad?

```
> prop.test(x = c(vaccine_gotflu, placebo_gotflu),
+           n = c(vaccine_samplesize, placebo_samplesize))

	2-sample test for equality of proportions with continuity correction

data:  c(vaccine_gotflu, placebo_gotflu) out of c(vaccine_samplesize,
placebo_samplesize)
X-squared = 59.891, df = 1, p-value = 1.003e-14
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.11436416 -0.06949298
sample estimates:
    prop 1     prop 2 
0.02057143 0.11250000 
```

In a medical trial for a new flu vaccine, is the proportion of participants who received the vaccine and then got the flu significantly different from the proportion who received a placebo and then got the flu?

Linear Regression

- Today: Bivariate (aka Simple) Linear regression, with two **numerical variables**: X (independent variable) and Y (dependent variable).

- Goal: prediction. What's our **best guess** of Y_i ("the value of Y for observation i") if we know X_i ("the value of X for observation i")?

- Simplest possible way to relate two variables: **a line**. You may remember from school the equation for a line: $y = mx + n$.

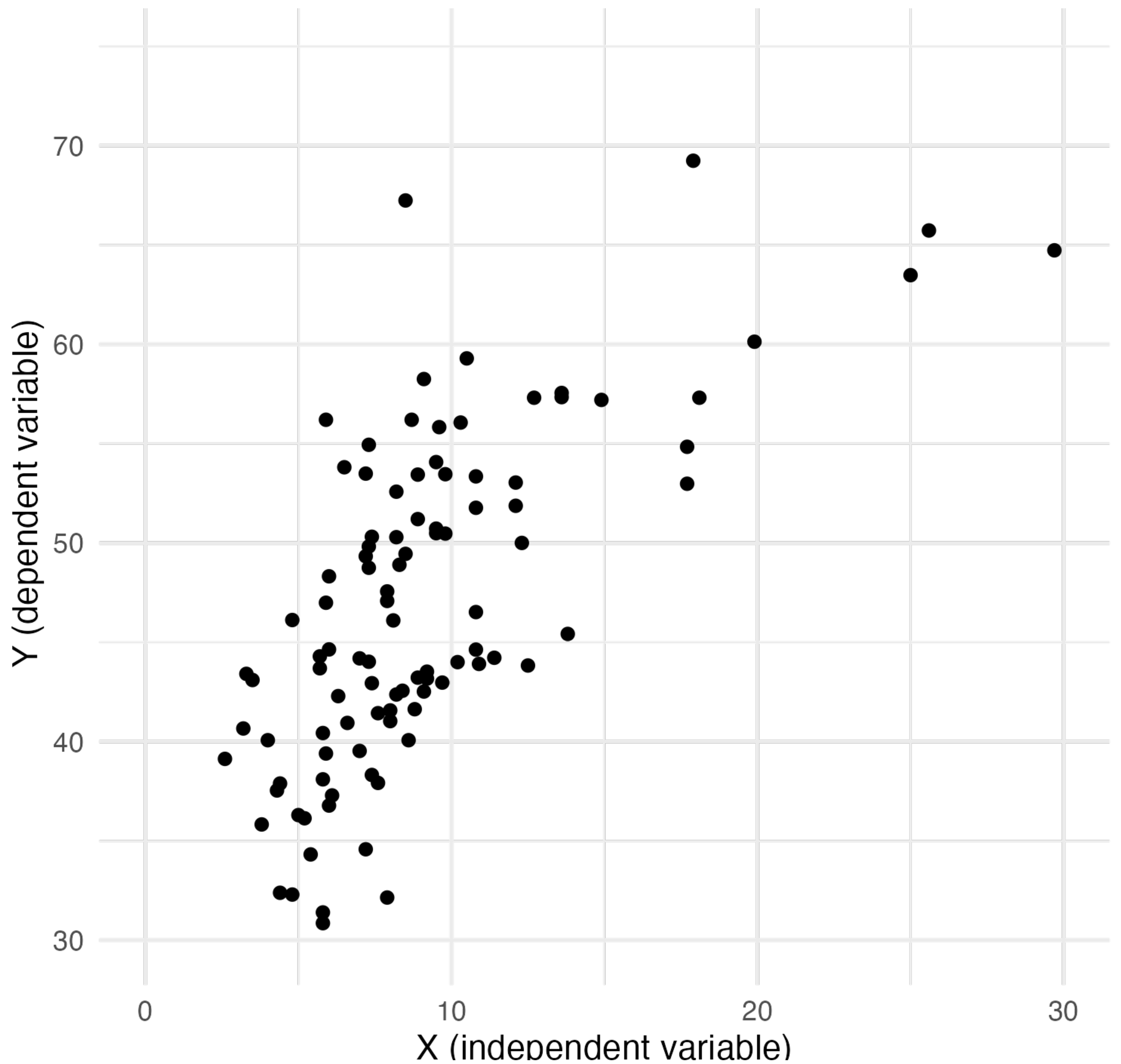
- Same here, but with Greek letters and indexes (optional): $Y_i = \alpha + \beta X_i$

Linear Regression

$$Y_i = \alpha + \beta X_i$$

- The problem: not all the data is going to be on the line. The world is messy.
- For any 'sensible' line we draw, some values of Y will be above the line, others below the line.
- So we model the line with some error ε, which may differ for each observation:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$



Linear Regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- This is a **model** of the process that generates Y: Y is a function of X plus some random error. It's a mathematical representation of reality.
- α is the **intercept** parameter: the predicted value of Y when $X = 0$.
- β is the **slope** parameter: the predicted change in Y associated with a one-unit increase in X. The slope is usually what we're most interested in.
- This is a **linear** model: by assumption, for ***every*** one-unit increase in X, we will see a corresponding increase in Y by β amount.

universität wien

Linear Regression

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

- α and β are parameters in our model: unknown features of the data-generating process, which we do not directly observe because our data comes with some random error.

- We actually **estimate** $\hat{\alpha}$ and $\hat{\beta}$ from our observed data, by figuring out the “best” line to fit through our data:

Predicted (or “fitted”) values of Y → $\hat{Y}_i$

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

Estimates for the intercept and slope → $\hat{\alpha}$, $\hat{\beta}$

Linear Regression

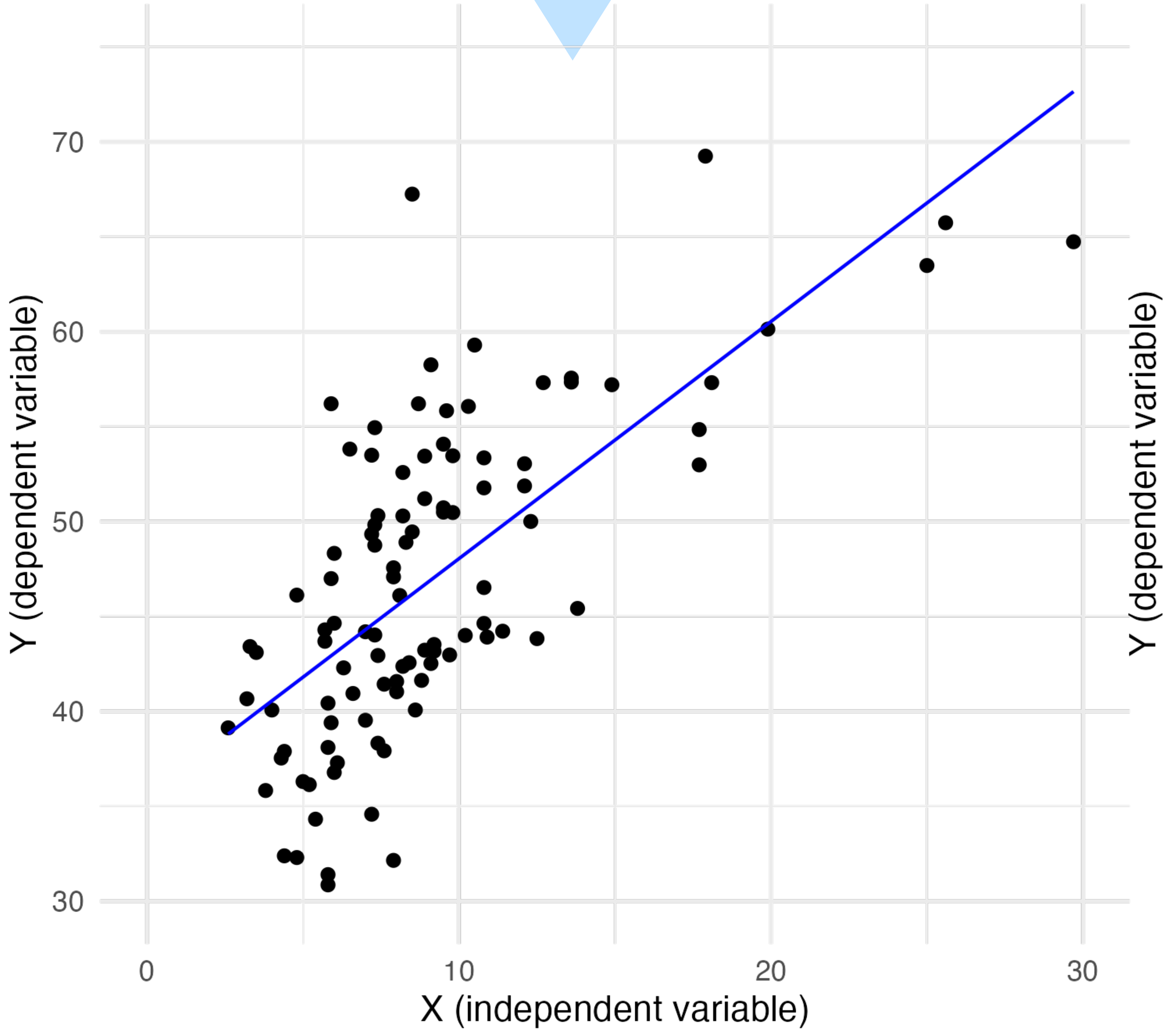
- How do we pick the "best" line $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$?
- Ordinary Least Squares: we choose $\hat{\alpha}$ and $\hat{\beta}$ so that they minimise the **sum of squared residuals**, where the residuals $\hat{\varepsilon}$ are the difference between the predicted values of $\hat{Y}$ and the observed values Y:

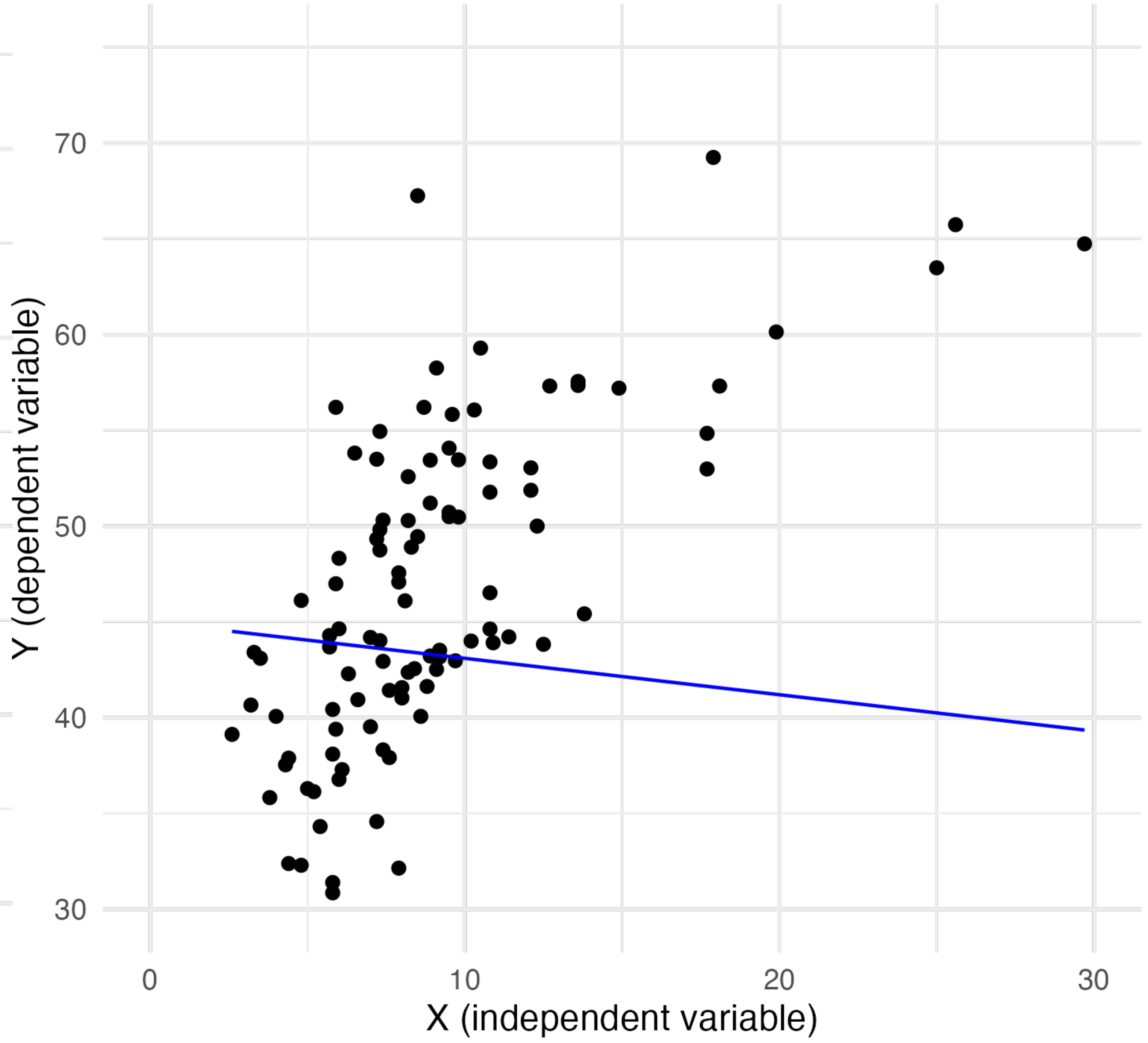
$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}X_i)^2$$

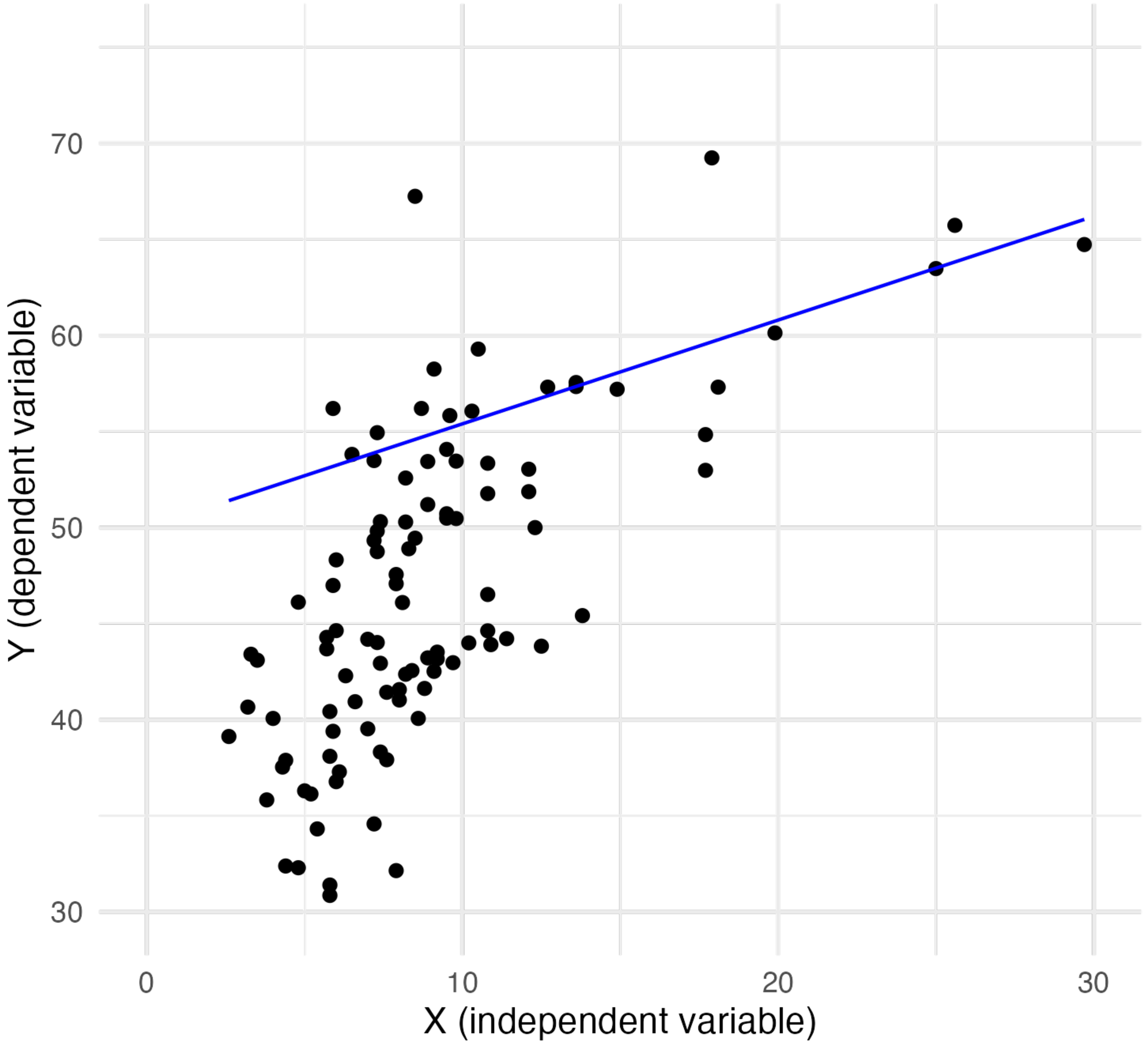
- You can solve for $\hat{\alpha}$ and $\hat{\beta}$ with calculus (but we'll let R do it for us!)

Visually...

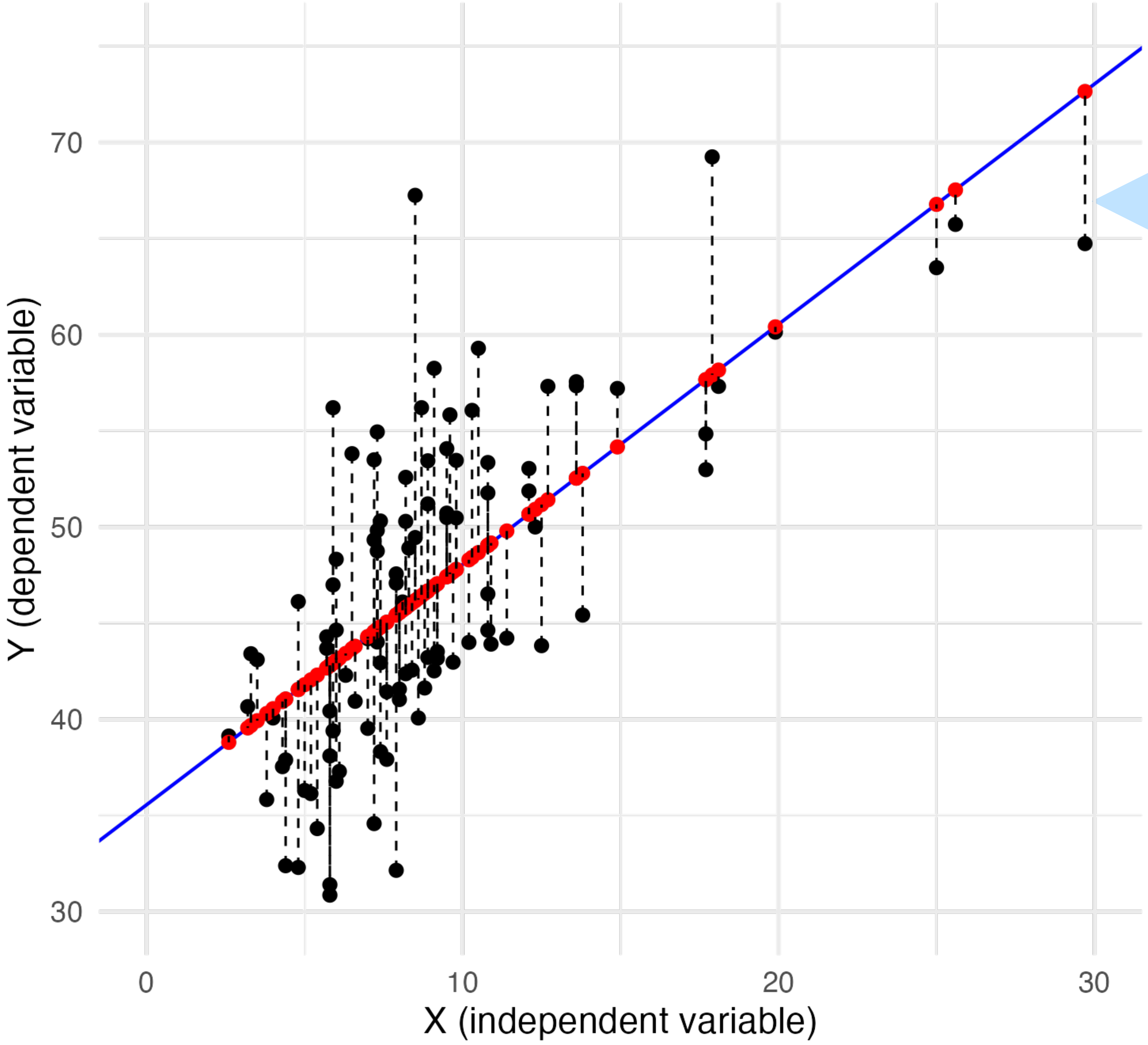
Of all possible lines, we pick this one...







Visually…

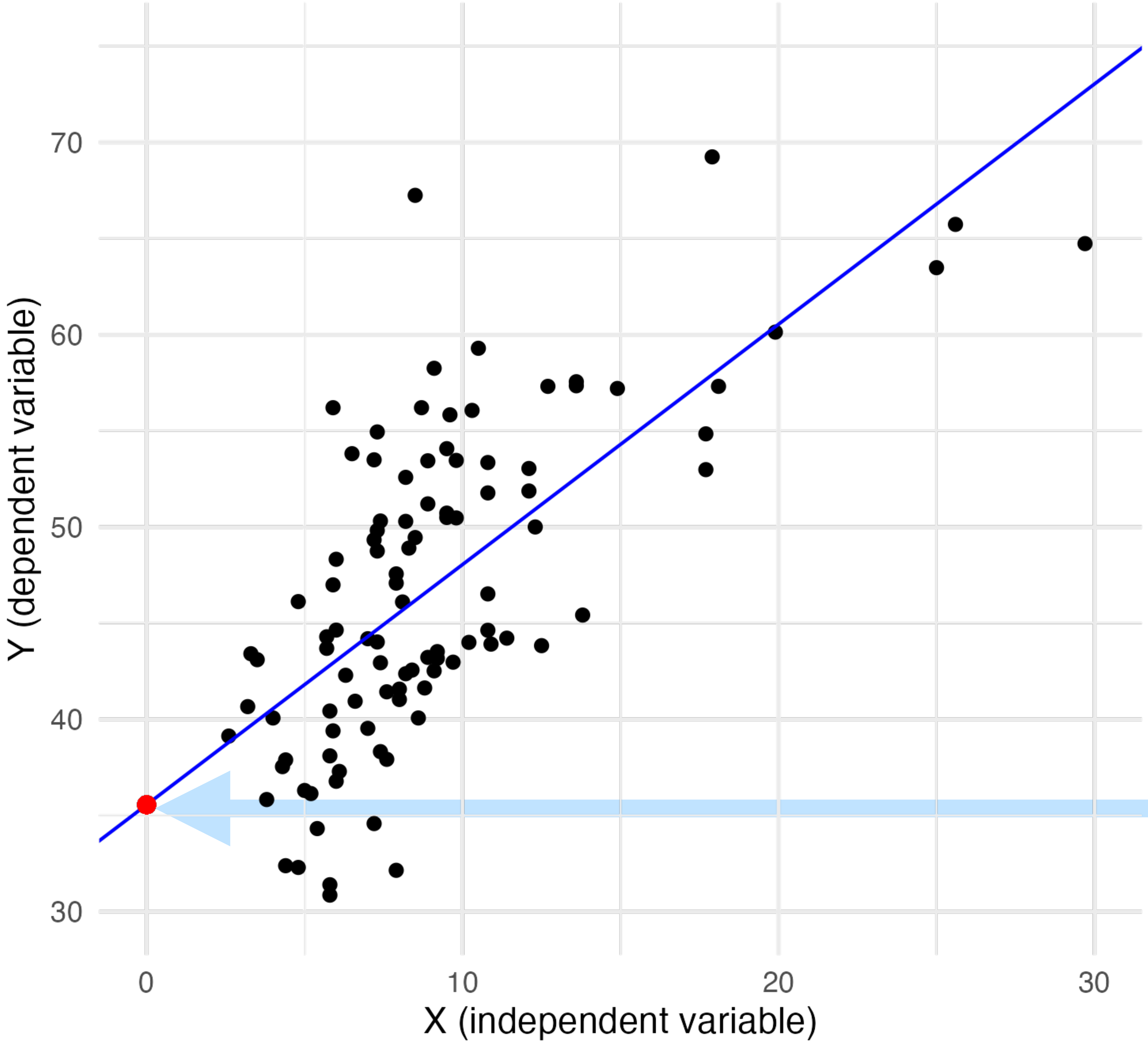


…because if we take all the difference between the observed values of Y (in black) and the predicted values $\hat{Y}$ (in red)…

…Then we **square** these differences (the residuals), so they all become positive…

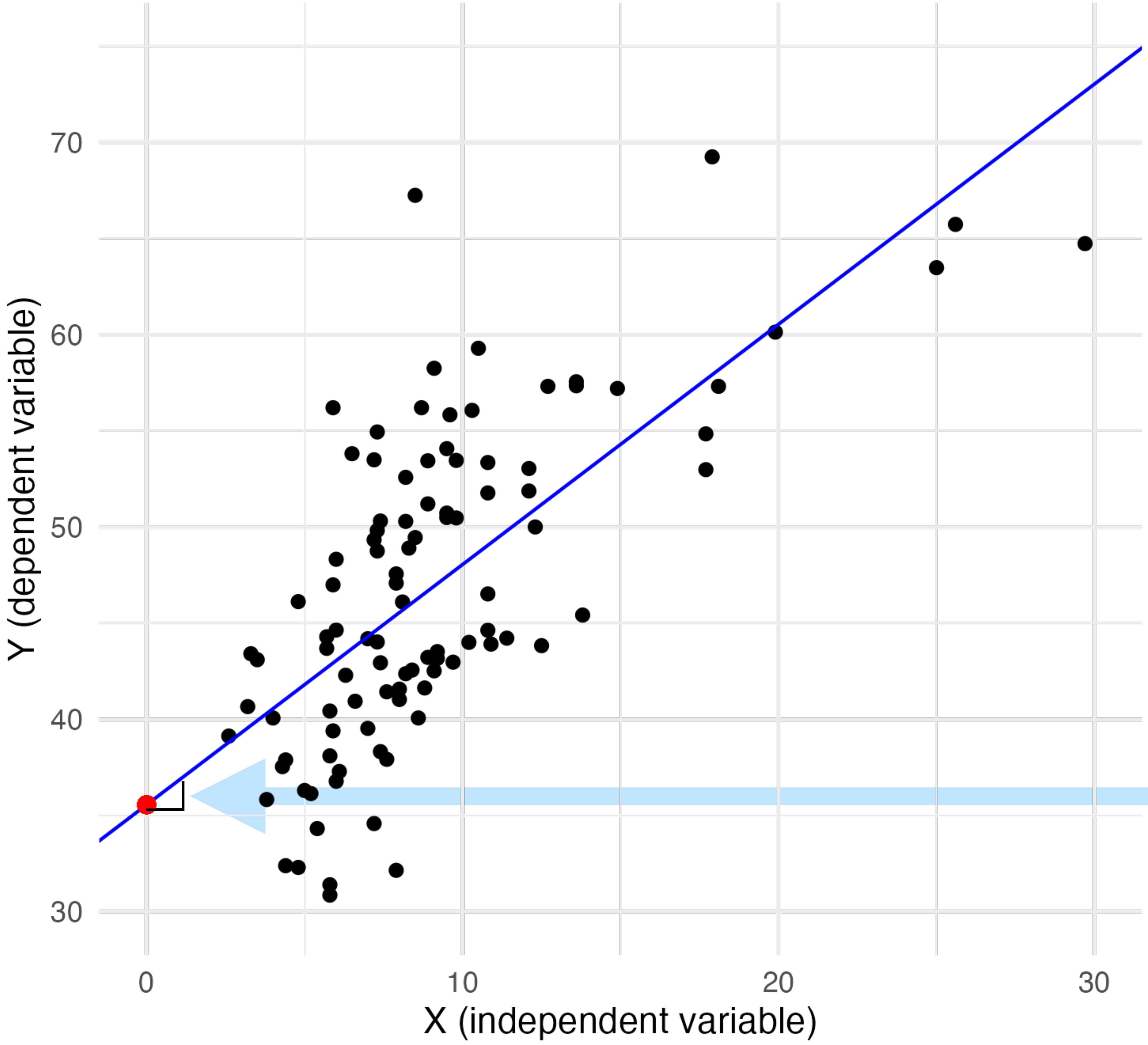
…and we sum them all up, this specific line returns the **minimum sum of squared residuals**.

Visually...



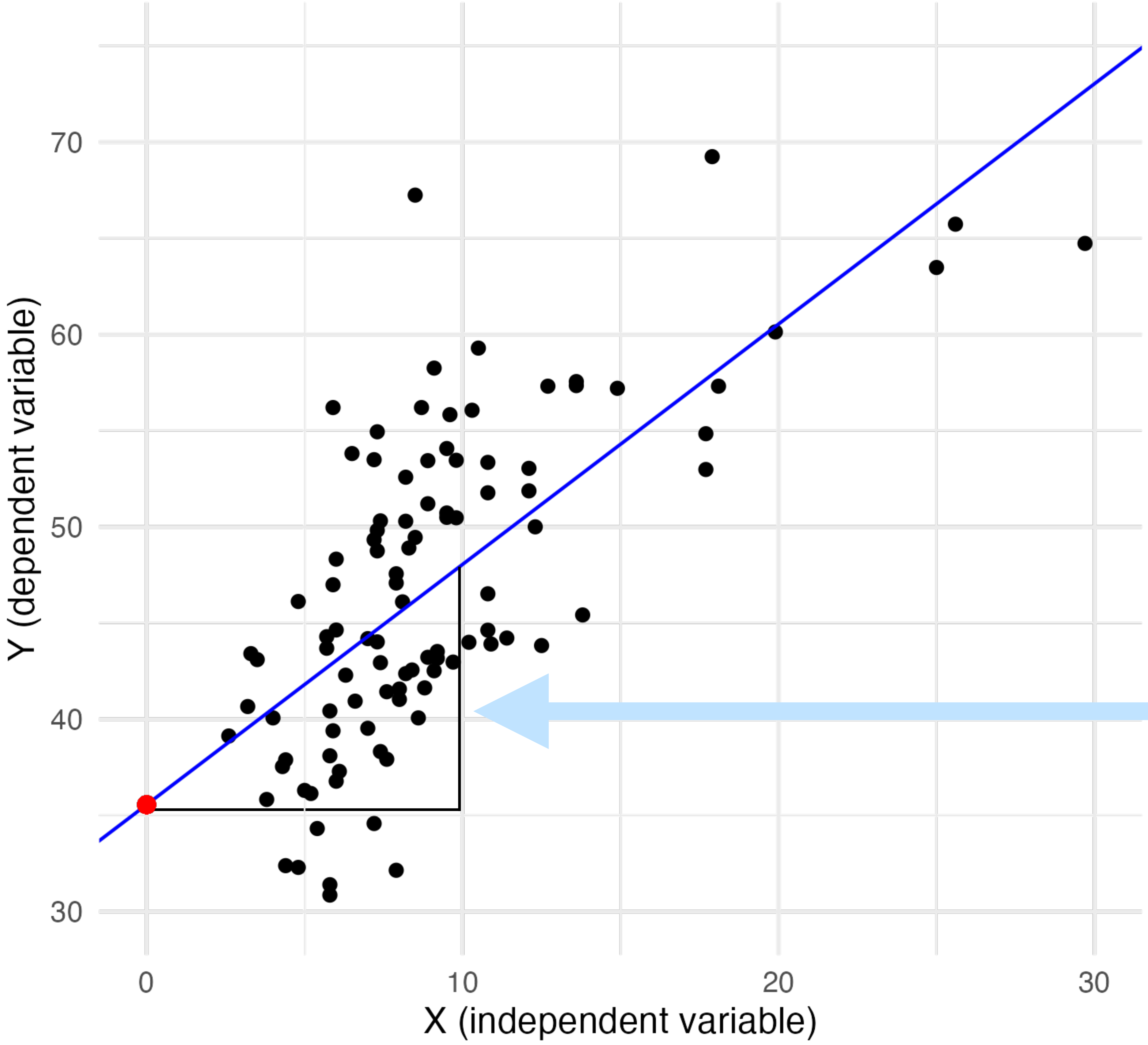
The intercept $\hat{\alpha}$ is the predicted value of Y when X is zero. In this case, about 35.

Visually...



The slope $\hat{\beta}$ is the predicted change in Y as we increase X by 1. In this case, it's 1.25

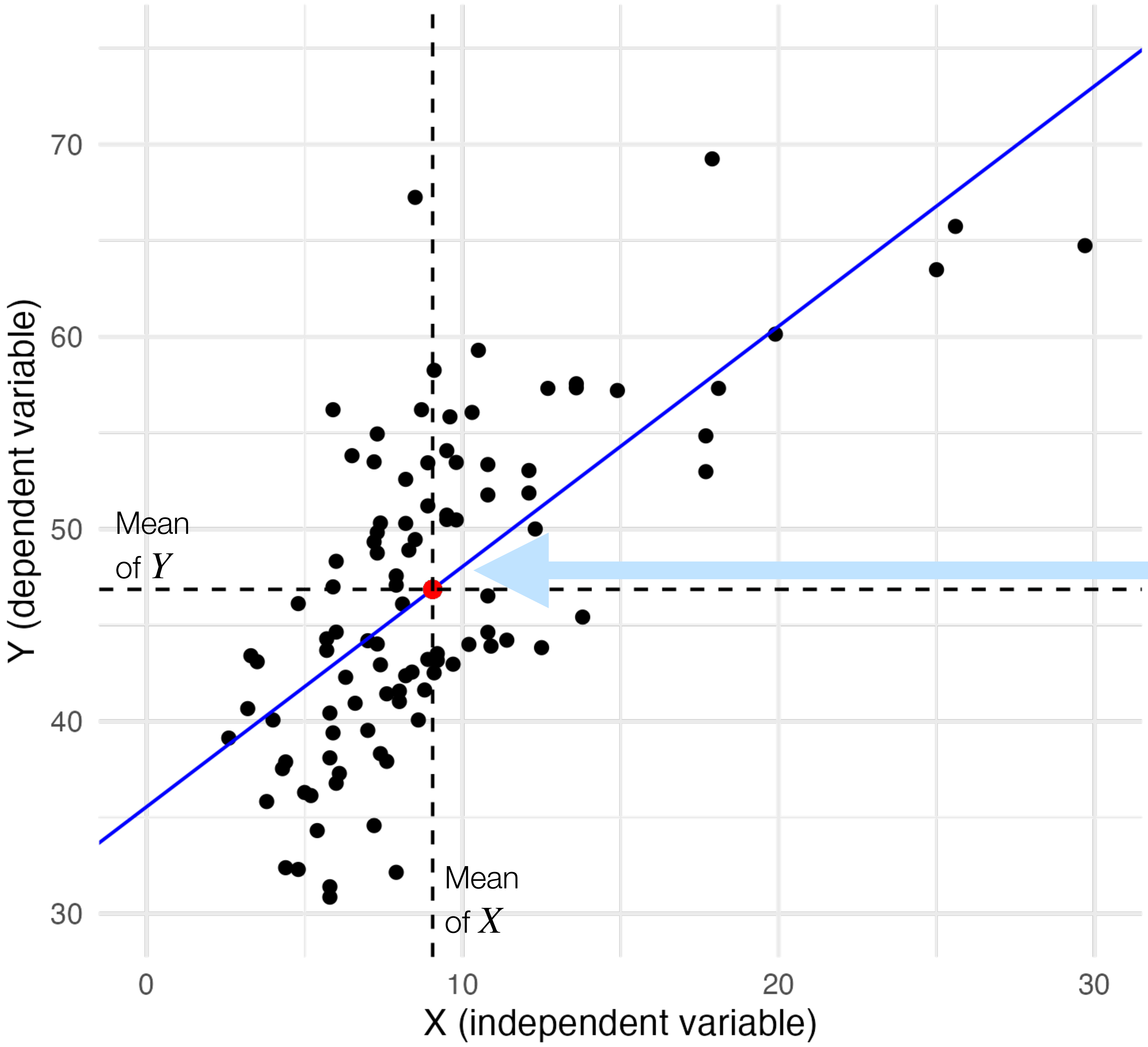
Visually…



Because we're fitting a line, the predicted increase in Y associated with a one-unit increase in X is the same ($\hat{\beta}$ = 1.25) everywhere…

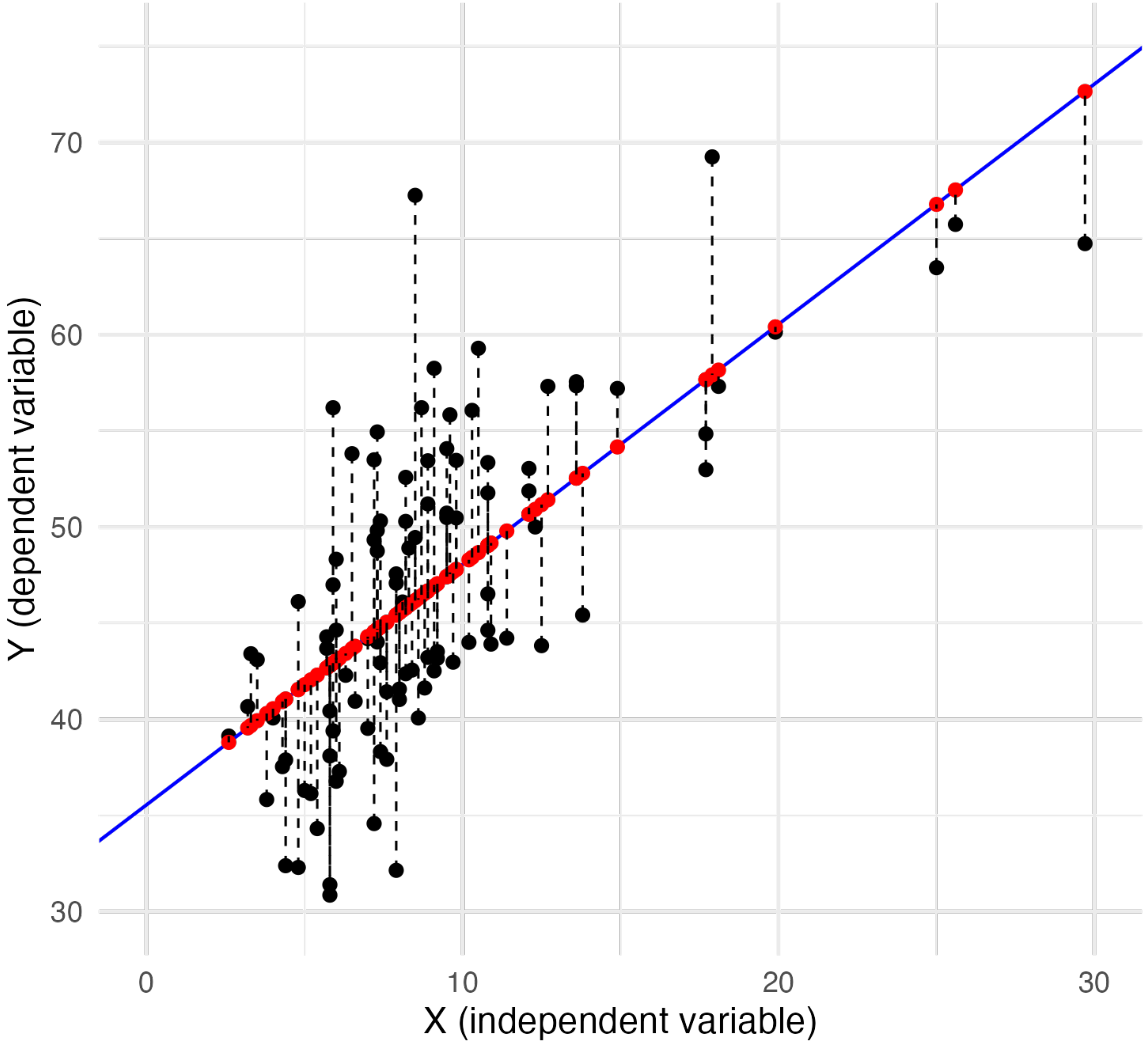
…so, for instance, increasing X by 10 is associated with a 12.5 increase in Y.

Visually…



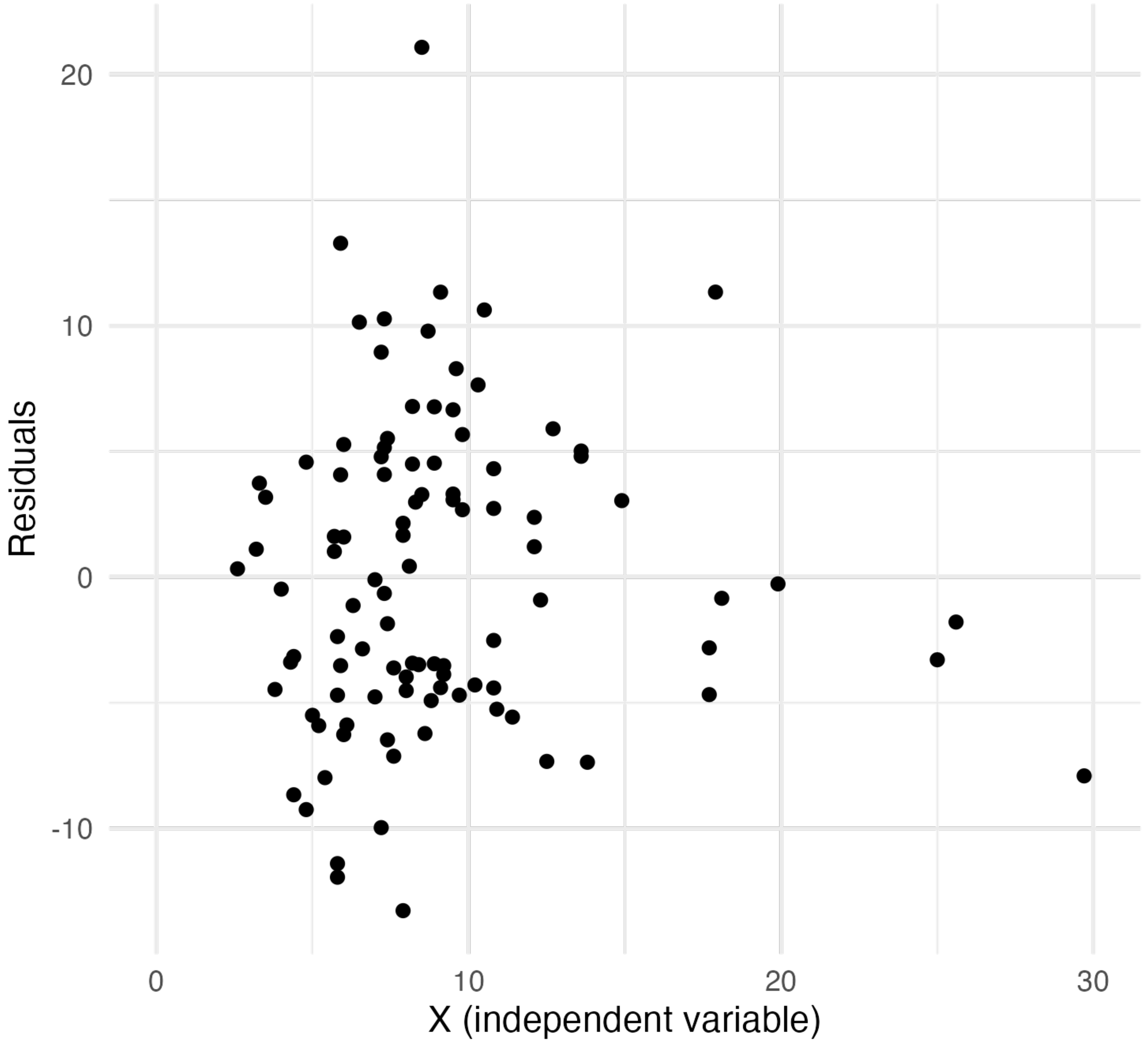
The regression line always runs through the mean of X and the mean of Y

Visually…



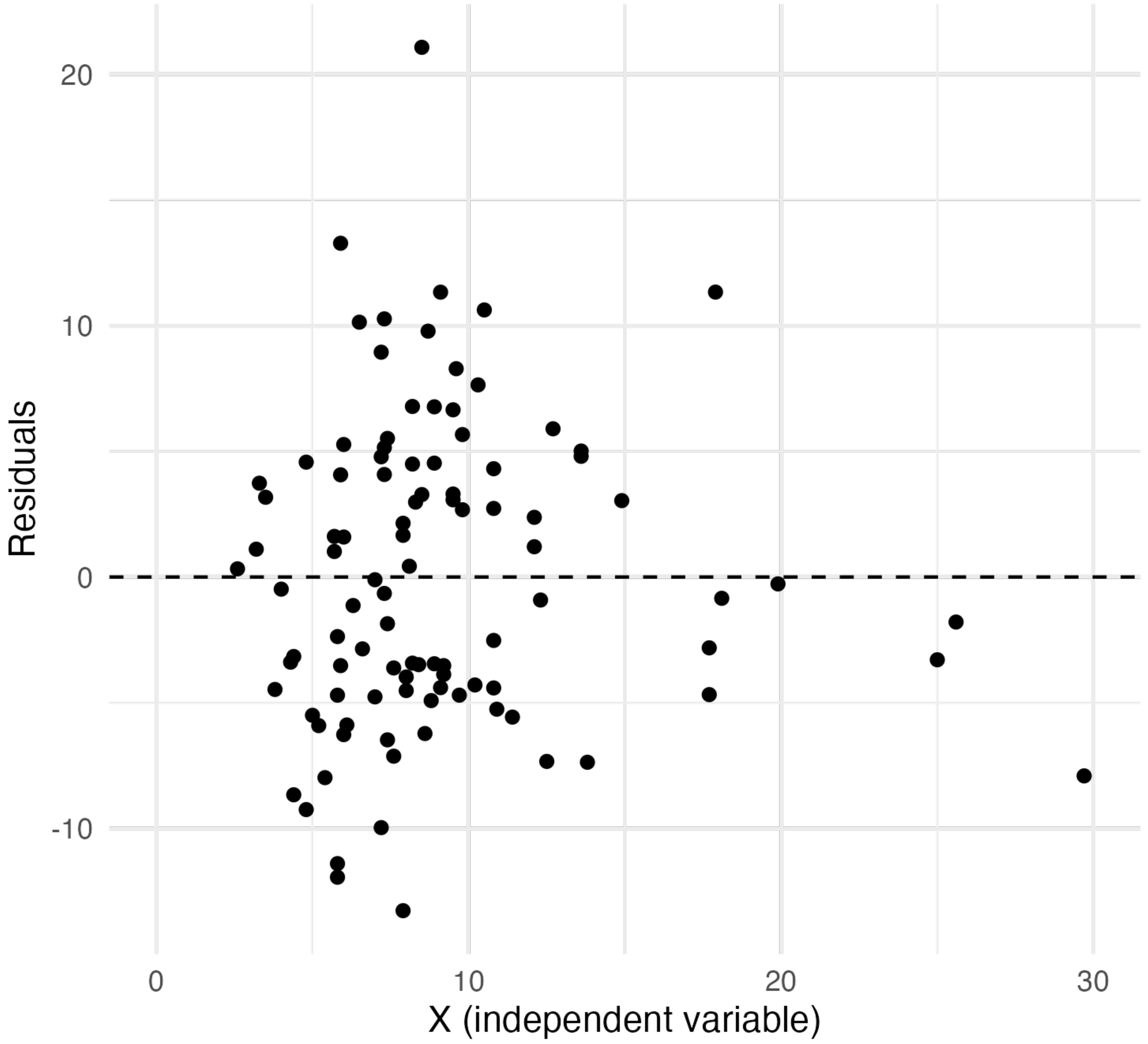
If we take all the residuals $(Y_i - \hat{Y}_i)$…

Visually…



And we plot them…

Visually…



Their mean will always be zero.

Linear Regression in R

- Example: Brexit data from earlier weeks. What's the expected change in "Leave vote" in a local authority associated with a one-unit increase in "Percentage of residents with a university degree"?

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    81.38653    1.24070   65.60   <2e-16 ***
percent_degree -1.04909    0.04432  -23.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$\hat{\alpha}$ (Intercept)

$\hat{\beta}$ percent_degree

Linear Regression in R

- Example: What's the predicted change in life satisfaction (0-10 scale) associated with a one-unit increase in religiosity (0-10 scale)?

$\hat{\alpha}$

$\hat{\beta}$

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    6.5526     0.2173  30.150   <2e-16 ***
## religiosity    0.1053     0.0471   2.236   0.0262 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

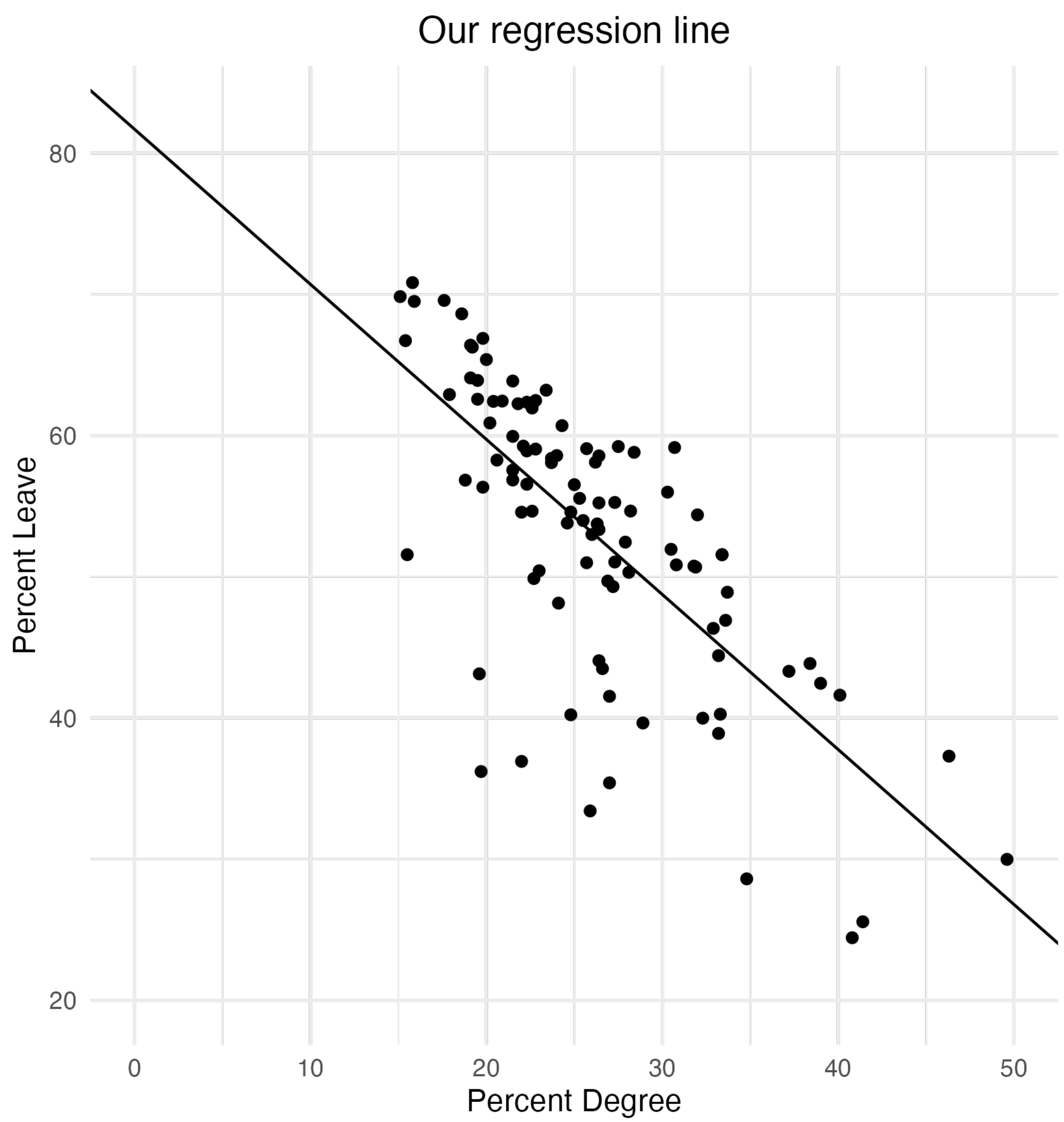
Goodness of fit (R^2)

- Normally, the thing we're most interested in when we fit a regression is the slope coefficient.
- Interpretation: "$\hat{\beta}$ represents the predicted change in Y associated with a one-unit increase in X".
- It comes with its measures of uncertainty (week 13) and under very restrictive assumptions, it may be interpreted as an **effect** (week 14).
- However, we may also be interested in finding out how well our linear model explains variation in Y.

Goodness of fit (R^2)

- The measure of "goodness of fit" is the R^2.
- Suppose we have fitted our regression line for this model:

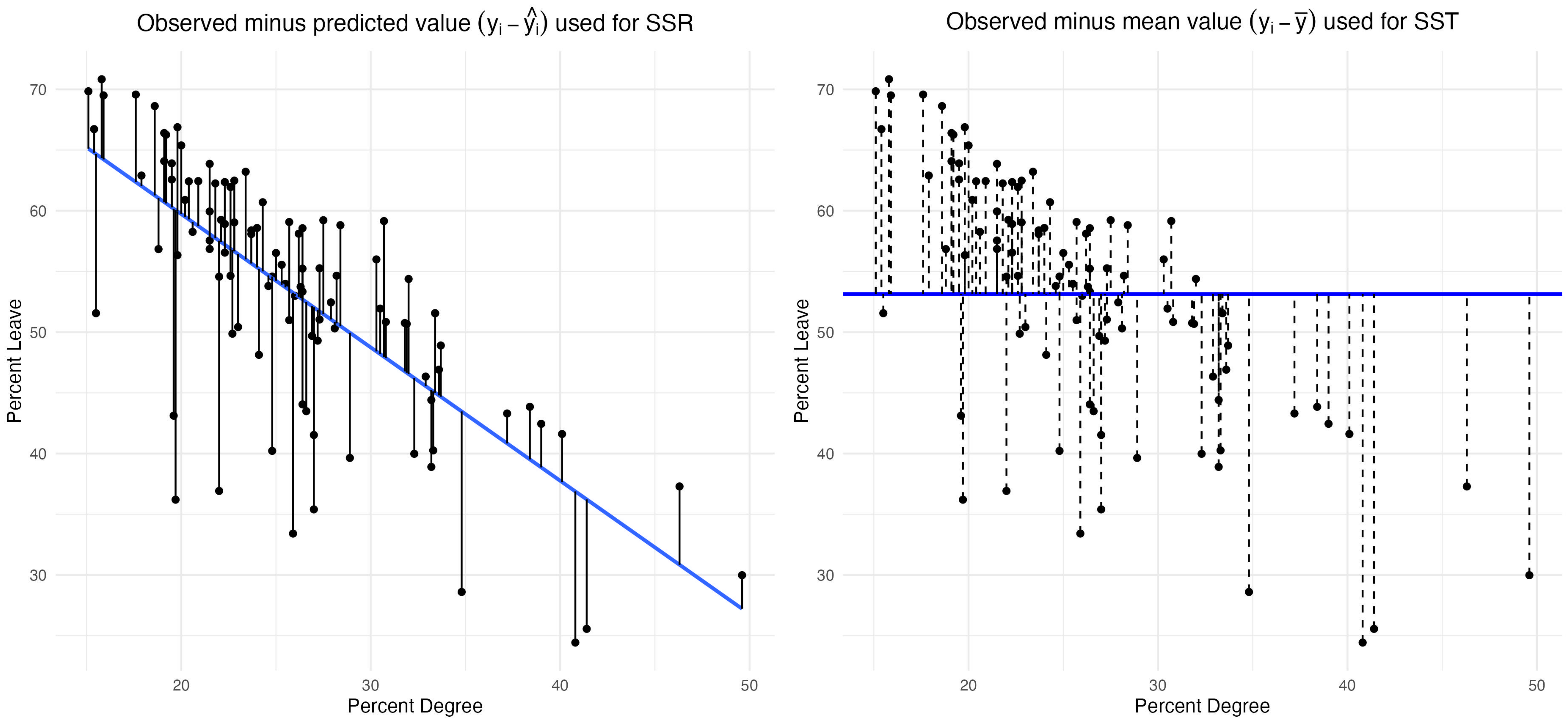
$$\text{Leave Vote}_i = \alpha + \beta(\text{Pct. Degrees}_i) + \varepsilon_i$$



Goodness of fit (R^2)

- Logic of the R^2:
- Compare the unexplained variation in Y ***after*** fitting the line with the unexplained variation in Y ***before*** fitting the line
- Before fitting the line: best guess for any value of Y is the mean $\bar{Y}$.
- After fitting the line: best guess for Y_i is the predicted value $\hat{Y}_i$.

Goodness of fit (R^2)



Goodness of fit (R^2)

- The R^2 compares the Sum of Squared Residuals (unexplained variation after fitting the line, SSR), with the Sum of Total Squared (unexplained variation before fitting the line, SST)

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

- Interpretation: “the model explain $R^2 \times 100\,\%$ of the variance in Y.

R^2 in R

```
Call:
lm(formula = percent_leave ~ percent_degree, data = brexit)

Residuals:
    Min      1Q  Median      3Q     Max 
-26.067  -1.911   1.724   4.345  15.082 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    81.38653    1.24070   65.60   <2e-16 ***
percent_degree -1.04909    0.04432  -23.67   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.62 on 378 degrees of freedom
Multiple R-squared:  0.5971, Adjusted R-squared:  0.5961 
F-statistic: 560.3 on 1 and 378 DF,  p-value: < 2.2e-16
```

R^2

Good to know…

- Connection with other measures of correlation (from week 6)
- In a **bivariate** (simple) linear regression…
  - …the R^2 is the square of Pearson's r…
  - …and the slope coefficient $\hat{\beta} = \frac{\text{Cov(X, Y)}}{\text{Var(X)}}$
- But linear regression is more widely used because it can go beyond describing bivariate relationships: it can give us predictions for Y as a function of ***more than one*** X variable (next week).

universität wien

Summing Up…

- Linear regression allows us to make **predictions** about a dependent variable Y, based on the known values of the independent variable X.
- Line of best fit minimises the "sum of squared residuals". It is described by two values: the intercept ($\hat{\alpha}$) and slope ($\hat{\beta}$) coefficients.
- We're normally interested in interpreting $\hat{\beta}$: it's **the predicted change in Y associated with a one-unit increase in X.**
- We also get the R^2: it's the percentage of variance in Y explained by the linear model. **Your goal in life is not to maximise the R^2.**