THE SIMPLE REGRESSION MODEL I

PROF. SEBASTIÁN VALLEJO VERA

- For the first two weeks, I tried to convince you that experiments were the shit, allowing us to make simple comparisons that we could ultimately deem causal.
- I also showed you that, through the creative power of math, we could model our comparisons as a regression. I could see your eyes glistening with anticipation, craving knowledge.

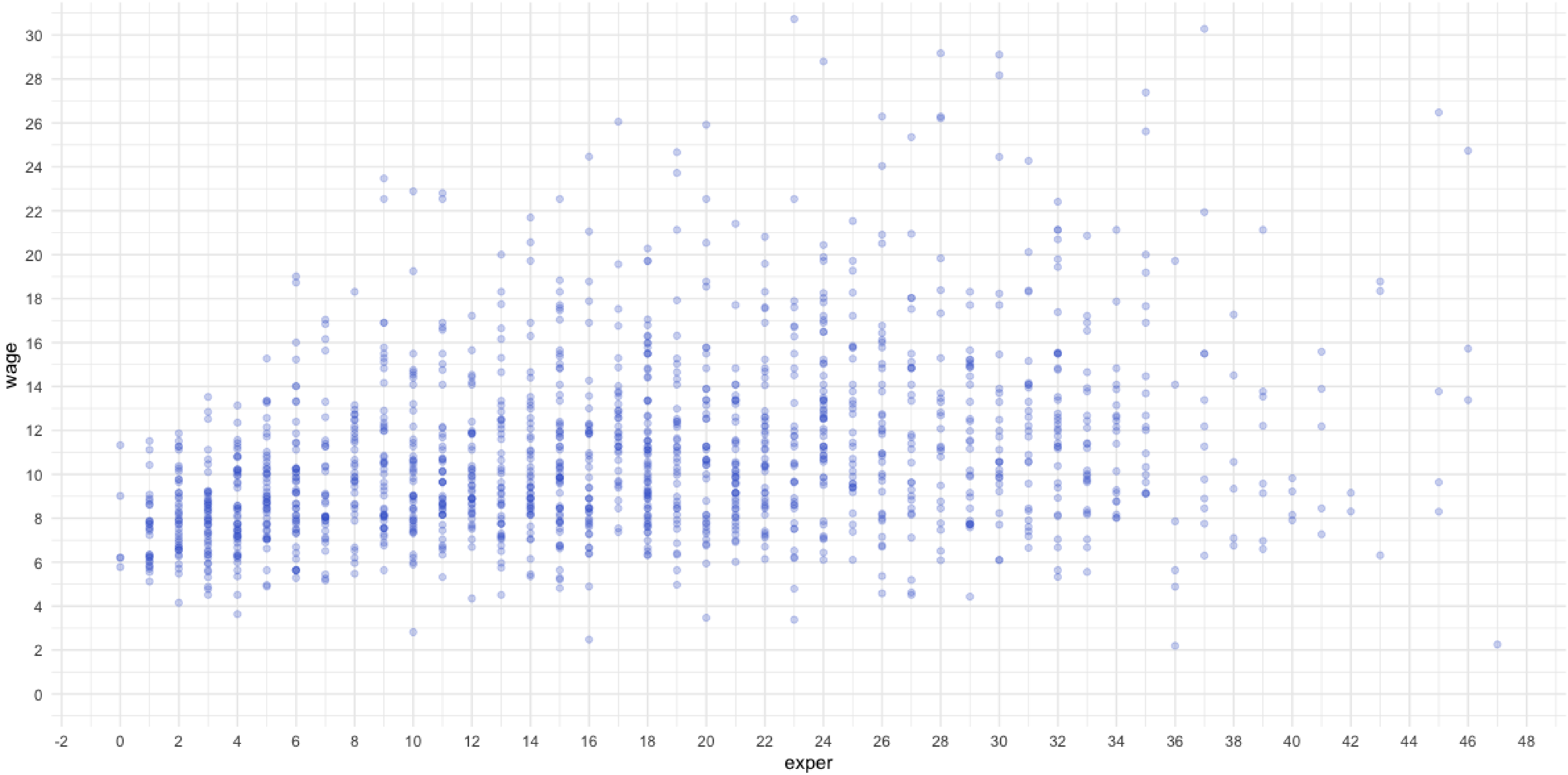
- But… why regressions? Didn't we just establish that with experiments we solve the fundamental problem of causal inference, we eliminate selections bias, we compare apples to apples, we raise our self-esteem, etc.?

- Why can’t we solely rely on experiments to estimate causal relations?
- That’s right:
  - It is not always possible
  - It can be costly
  - It can be immoral
- Furthermore, when we are not able to run experiments, we will have to use other methods to estimate causality and those will require regressions.
- Thus…

REGRESSIONS

- Remember hypothesis testing? What was the goal?

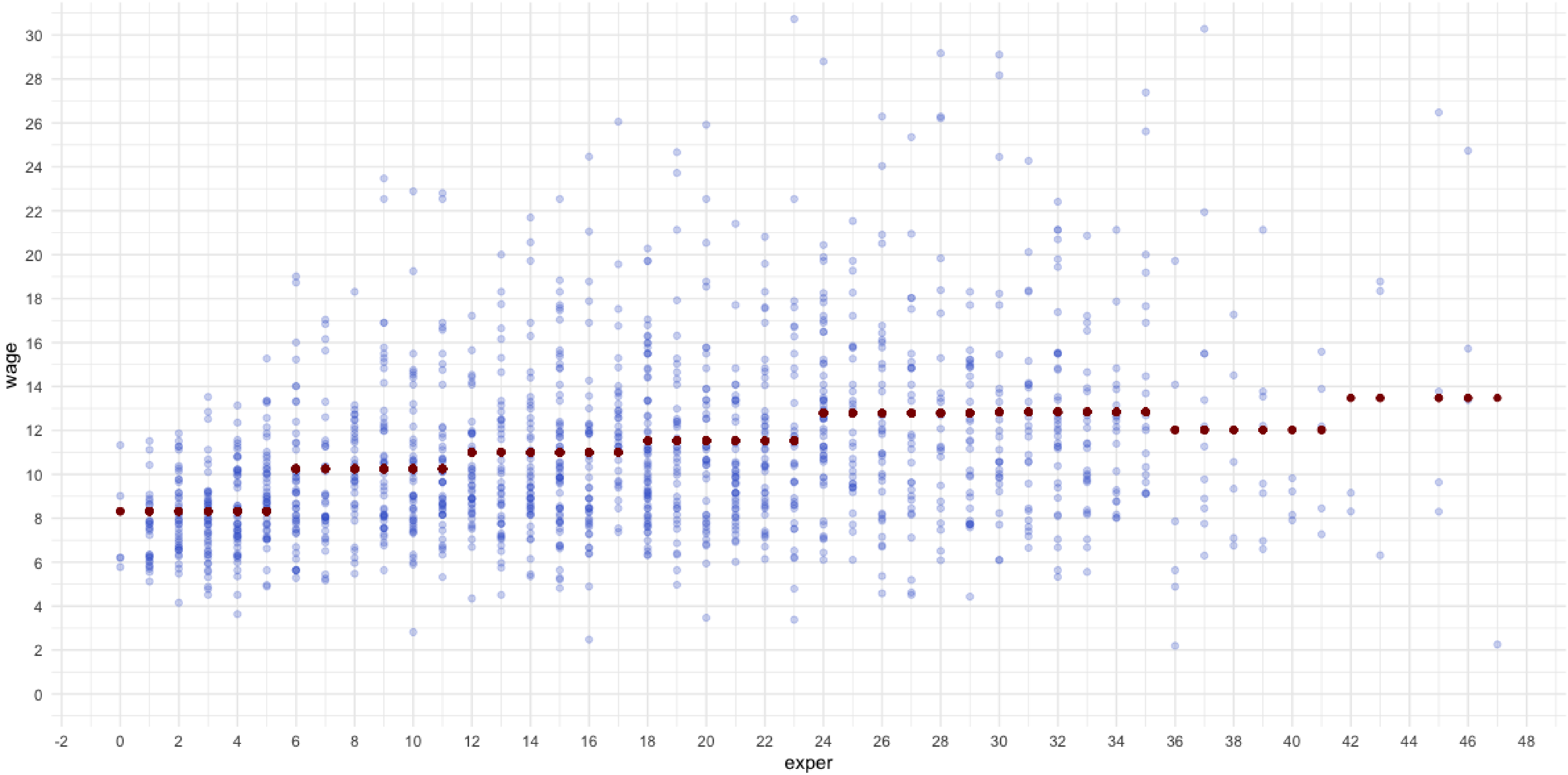
- Yes, to show how changes in one variable affect the changes in another.
- Let's say we have some research question about experience and wages, and we want to test the effect of experience on wages.
- Here is the data:



wage
exper

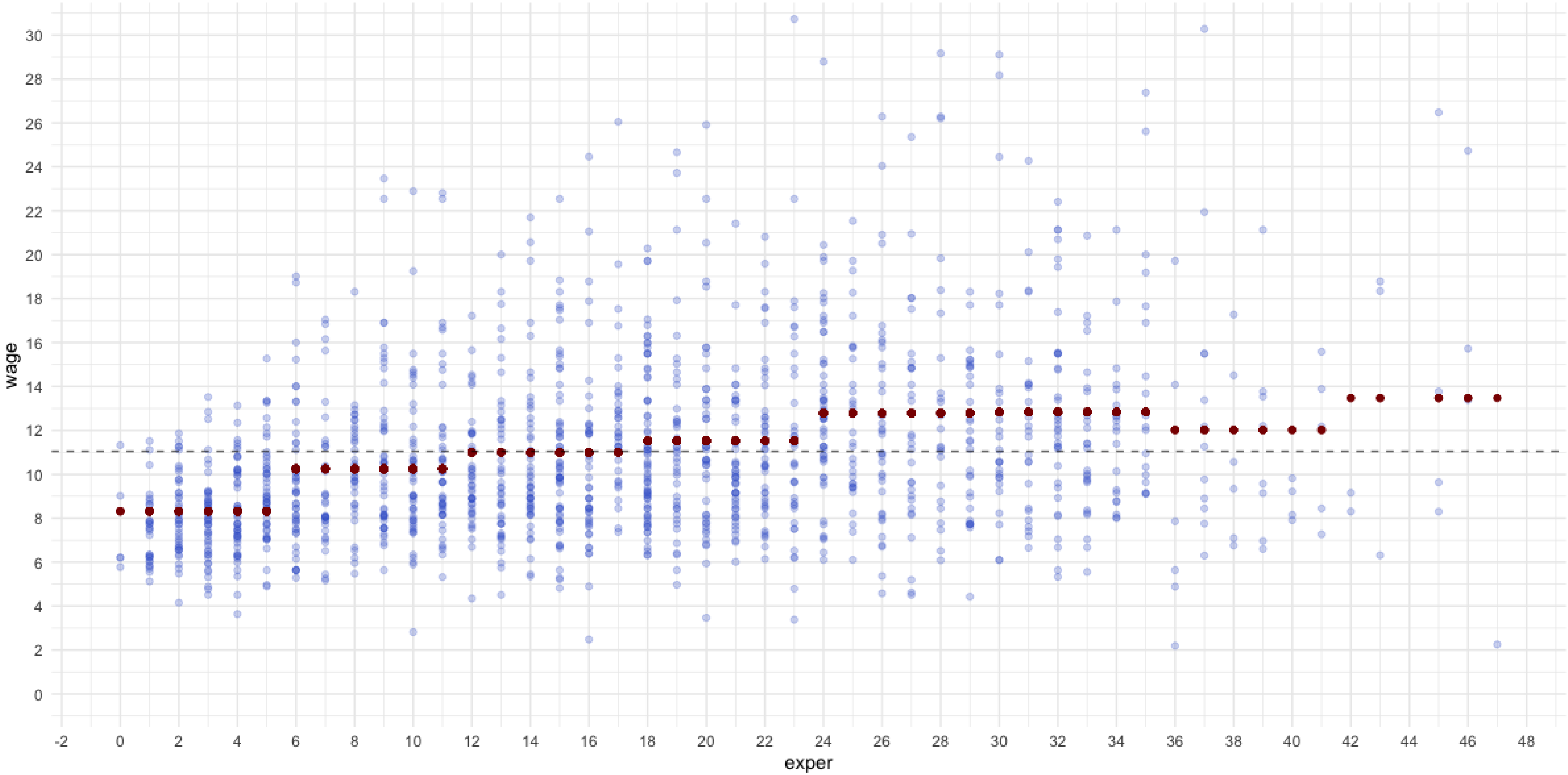
- How would you characterize this relationship (if you knew nothing about regressions)?

- One (intuitive?) way of doing this would be to divide the treatment (i.e., independent variable, x) into groups or bins, and see if applying different magnitudes of that treatment changes, *on average*, the outcome (i.e., dependent variable, y).



wage
exper

- There seems to be a positive relation: as experience increases, so do wages.
- Another way to understand this is to compare A) how well changes in experience explain changes in wages to B) how well we can explain wages by the tendency of normally distributed variables to gather around the mean (i.e., how well I can explain the likelihood of your wage by how far you are from the mean).


wage
exper
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
-2
32
34
36
38
40
42
44
46
48

- Ok, so, what’s the difference between my grouping technique and a regression?
- In a regression, we can give more structure to our predictions.

- But before this, we need to make certain assumptions. The first assumption we are going to make is about our two variables of interest (x and y): that their relation is linear (i.e., that the values of y increase/decrease monotonically across the values of x)*

- Since we are projecting this relationship in a two-dimensional space, it might be useful to think of the algebraic formula for a straight line:
  - $y = b + mx.$
- What does y, x, b, m mean?

* Later in the course we will see what happens when we violate this, and other assumptions.

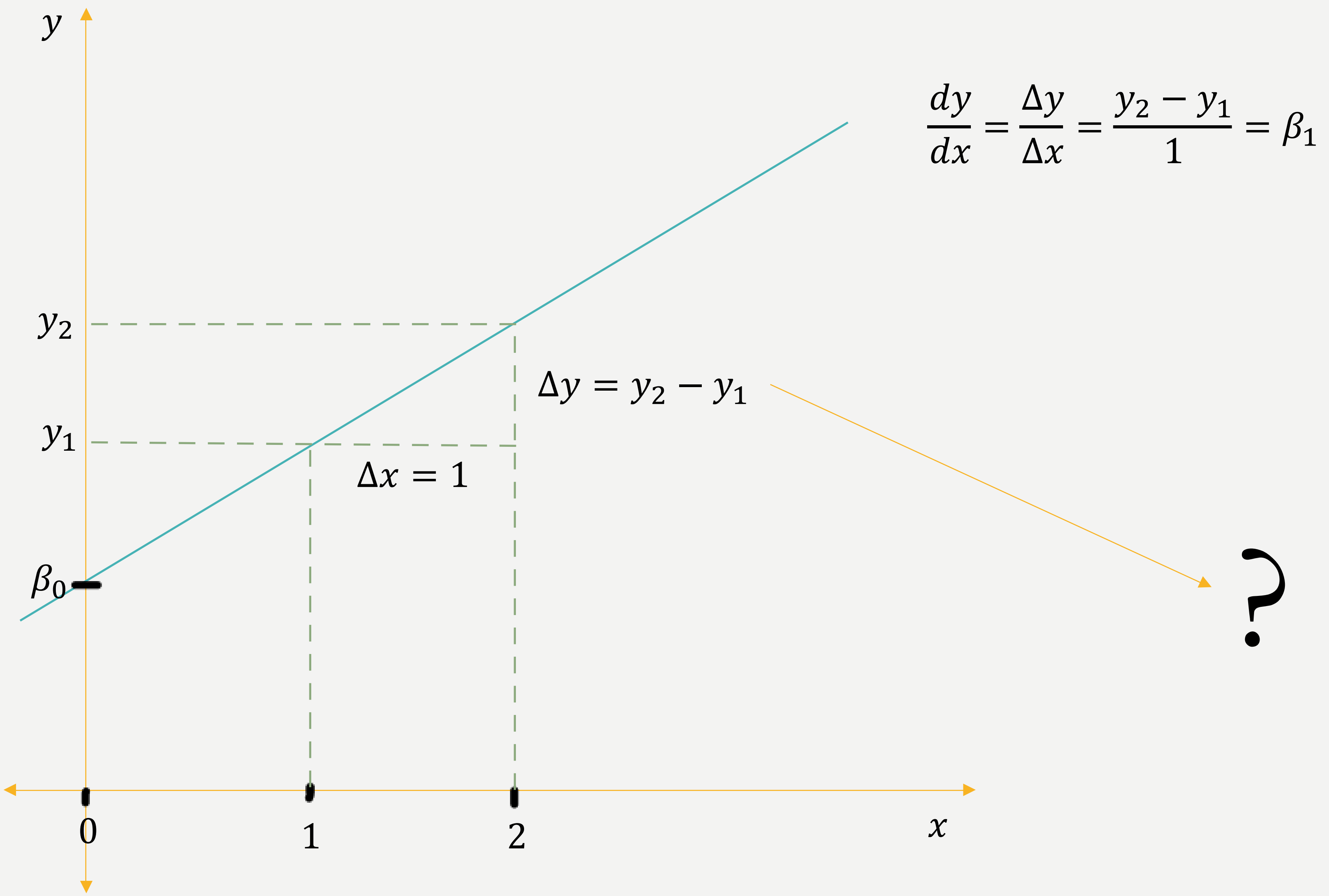
THE SIMPLE REGRESSION MODEL (SRM)

- Cool. Let's turn that into a **linear regression model**.
- To examine how y varies with changes in x, we propose a two-variable linear regression model (also known as a *bivariate linear regression model*) in the population of interest (the population regression function or PRF):
  - $y = \beta_0 + \beta_1 x + \mu$
- $\boldsymbol{y}$ is the dependent variable (or the explained or response or outcome variable, *wage*)
- $\boldsymbol{x}$ is the independent variable (or explanatory or control variable or covariate, *experience*).
- $\boldsymbol{\mu}$, called the error term or disturbance, represents factors other than x that affect y, the "unobserved."

INTERPRETATION

- Intercept (β_0): the value of y when $x = 0$. In other words, if my experience was zero years, what would my salary be. In our regression equation:
  - $y = \beta_0 + \ \beta_1 0 + \mu$
  - $y = \beta_0 + \mu$
- Slope (β_1): the pace at which y changes in relation to a one-unit change in x*.
  - $\frac{dy}{dx} = \frac{\Delta y}{\Delta x} = \beta_1$
  - This is true if the the other factors in μ are held fixed ($\Delta\mu = 0$).

* Note also that the linearity of $y = \beta_0 + \ \beta_1 x + \mu$ implies that a one-unit change in x has the same effect on y, regardless of the initial value of x. This is unrealistic in many applications.



y
$\frac{dy}{dx} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{1} = \beta_1$
y_2
$\Delta y = y_2 - y_1$
y_1
$\Delta x = 1$
β_0
?
0
1
2
x

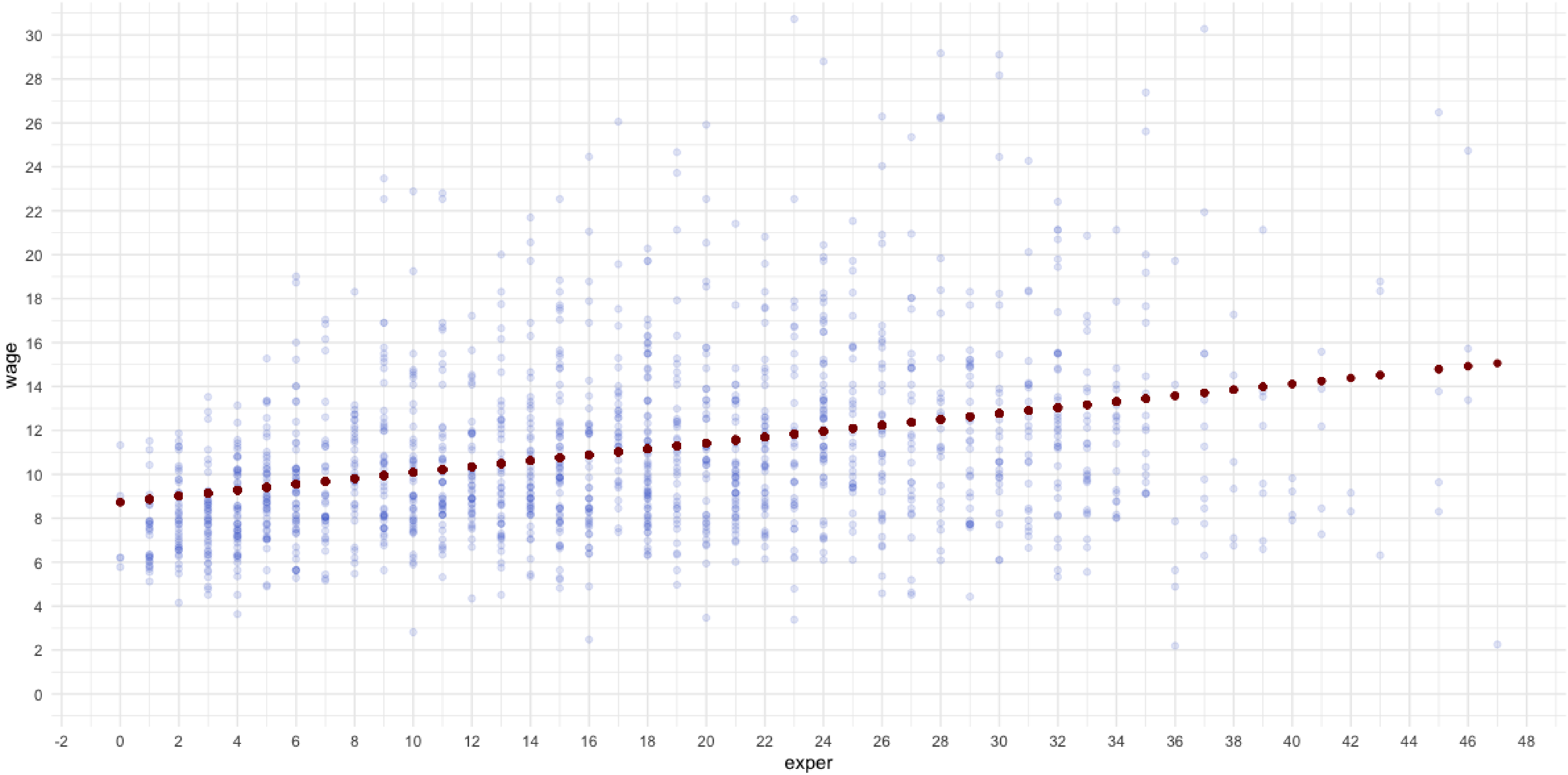
- Our interest lies in the estimation of the parameters β_0 and β_1.
- Recall that $y = \beta_0 + \ \beta_1 x + \mu$ is the population equation. This means that the relationship between x and y is real in the *population*, that is, it exists but it is unknown.
- But how do we go about estimating the relationship between x and y with a random sample from the population?

- Let $\{(x_i\ ,\ y_i): i = 1, \ldots, n\}$ denote a random sample of size n from the population. With these data in hand, we want to estimate the following equation:
- $\widehat{y_i} = \widehat{\beta_0}\ + \widehat{\beta_1} x_i$
- where $\widehat{\beta_0}$ and $\widehat{\beta_1}$ are estimated and $\widehat{y_i}$ becomes therefore predicted, as indicated by the hat (ˆ).
- There are several possible estimators to estimate $\widehat{y_i} = \widehat{\beta_0}\ + \widehat{\beta_1} x_i$ but the best linear unbiased estimator (under certain assumptions) is the Ordinary Least Squares (OLS). (More on that later)

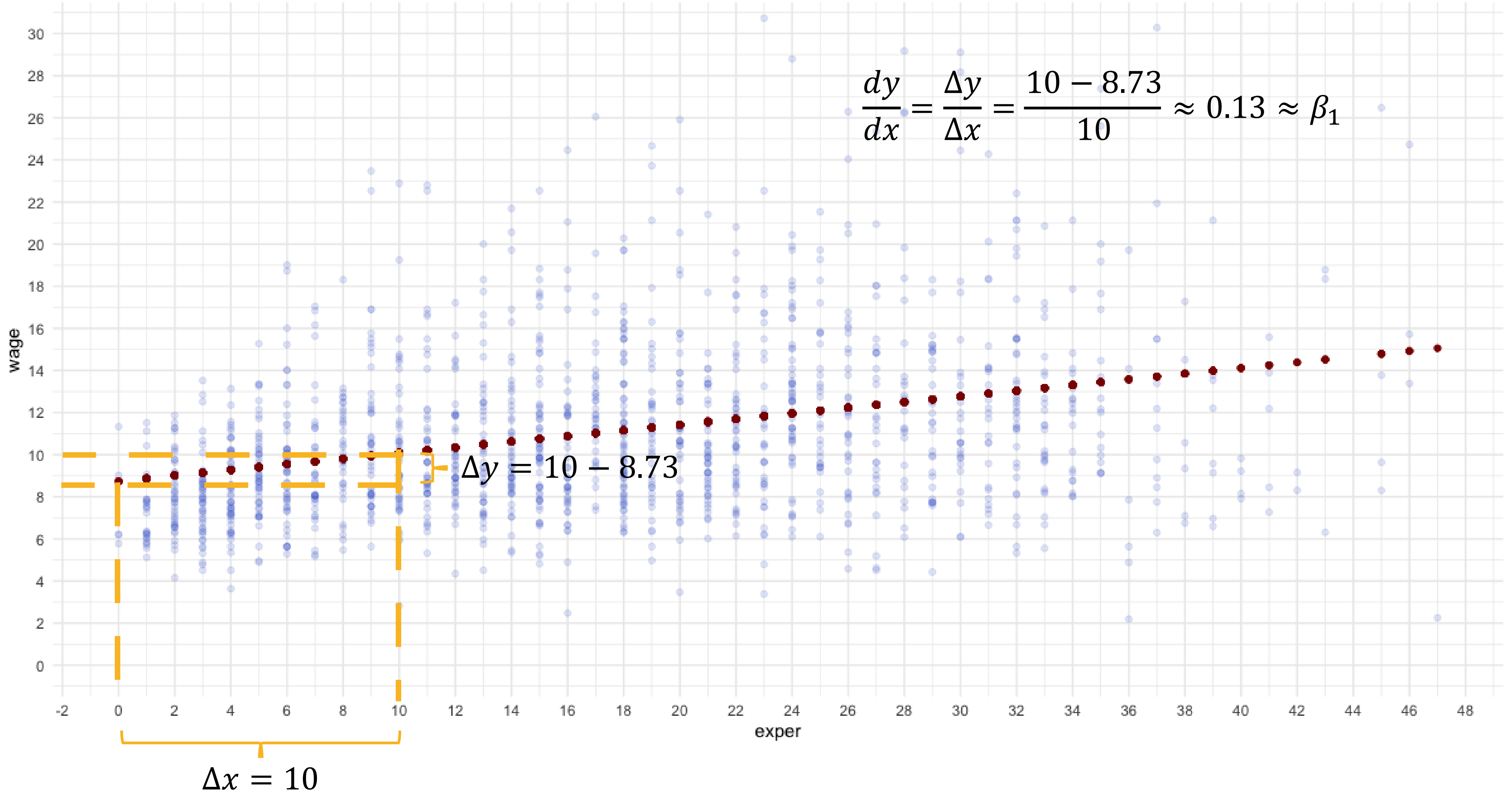
- Let’s think back about the relationship between experience and wages (in the population):
  - $wage = \beta_0 + \beta_1 experience\ +\ u$
- where *wages* is measured in euros per hour and *experiences* is years of experiences.
- By using the data in ‘Bwages’ from the ‘Ecdat’ package, we can plot experiences over wage and estimate a regression line.

```
> # Vignette 4.2: ----
>
> # What's inside a regression? An intercept, a slope... a line!!
> model <- lm(wage~exper,data=Bwages)
> summary(model)

Call:
lm(formula = wage ~ exper, data = Bwages)

Residuals:
    Min      1Q  Median      3Q     Max
-12.803  -2.554  -0.749   1.643  35.075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.73486    0.21723   40.21   <2e-16 ***
exper        0.13450    0.01087   12.38   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.237 on 1470 degrees of freedom
Multiple R-squared:  0.0944,	Adjusted R-squared:  0.09379
F-statistic: 153.2 on 1 and 1470 DF,  p-value: < 2.2e-16
```

$$\Longrightarrow \quad \widehat{wage} = 8.73 + 0.13\ experience$$

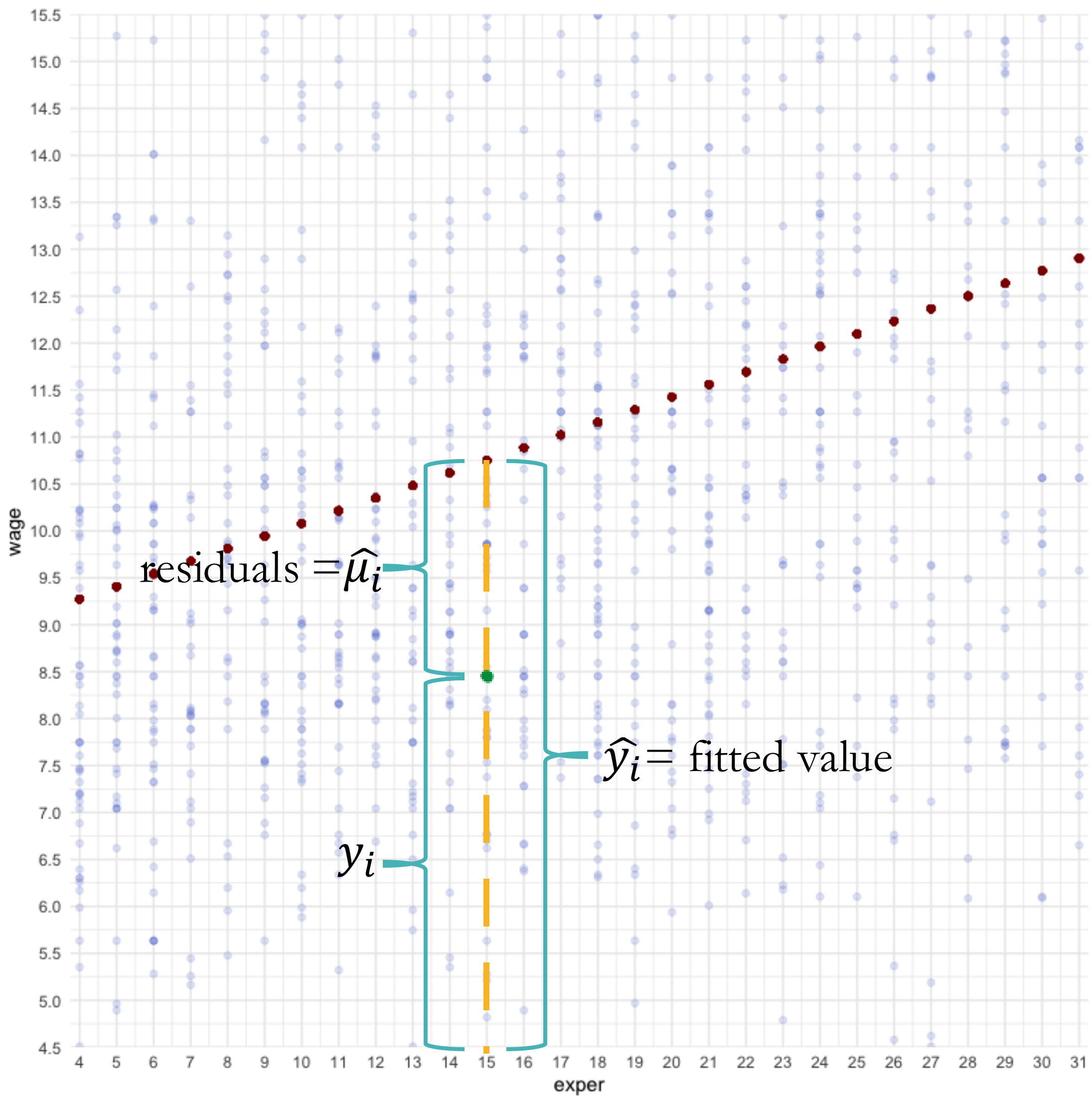


wage
exper
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
-2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48


$\frac{dy}{dx} = \frac{\Delta y}{\Delta x} = \frac{10 - 8.73}{10} \approx 0.13 \approx \beta_1$
$\Delta y = 10 - 8.73$
$\Delta x = 10$
wage
exper

- What does: $\widehat{wage} = 8.73 + 0.13\ experience$ tells us?
- First, the slope estimate $\widehat{\beta_1}$ tells us that one additional year of experience increases hourly wage by 13¢ an hour, on average.
- Therefore, 10 more years of experience increase the predicted wage by $10 * (0.13) = 1.30$, or $1.30 per hour, on average.
- In general, the intercept estimate $\widehat{\beta_0}$ is less meaningful (oftentimes, it's absurd), but in our case, it suggests that someone with no working experience would earn an hourly wage of $8.73, on average.
- Finally, it is possible to calculate predicted wages ($\widehat{wage}$), given different levels of experience. A person with 8 years of education would earn, on average, a wage of $9.77 per hour (in 1994 euros):
  - $8.73 + 0.13 * 8 = 9.77$

- Ok, Sebastián, that is neat trick, but… How does R know what is the line, among all the possible likes we could draw, that best fits our data?



$$\widehat{\mu_i} = y_i - \widehat{y_i}$$

$$\widehat{\mu_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$$

The best fit line is the line that minimizes the error. Ordinary Least Squares (OLS) comes from the fact that $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ are chosen to minimize the sum of the square residuals:

$$min \sum_{i=1}^{n} \widehat{\mu_i}^2 \longrightarrow \text{(i.e., RSS)}$$

$$\widehat{\beta_1} = \frac{cov(x_i, y_i)}{var(x_i)}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

…easy peasy

ALGEBRAIC PROPERTIES OF OLS STATISTICS

1. $\sum_{i=1}^{n} \widehat{\mu}_i = 0$
   - This is by construction, as the OLS estimates β_0 and β_1 were chosen to make the residuals add up to zero (for any dataset).
2. $\sum_{i=1}^{n} x_i \widehat{\mu}_i = 0$
   - This means that the sample covariance between the regressor x and the OLS residuals $\hat{\mu}$ is zero.
3. The point $(\bar{x}, \bar{y}\,)$ is always on the OLS regression line.

Note that because $\sum_{i=1}^{n} \widehat{\mu}_i = 0$, we have that $\bar{\widehat{y}}_i = \bar{y}_i$ since $\widehat{\mu}_i = y_i - \widehat{y}_i$

We can view OLS as decomposing each y_i into two parts, a fitted value and a residual:

1. $TSS \equiv \sum_{i=1}^{n}(y_i - \bar{y})^2$
   - TSS is the total sum of squares, and it measures the total sample variation in the y_i; that is, it measures how spread out the y_i are in the sample, or how much there is to explain.
2. $ESS \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
   - ESS is the explained sum of squares and, similarly, measures the sample variation in the $\hat{y}_i$; that is, how much of the variation is explained by the model.
3. $RSS \equiv \sum_{i=1}^{n} \hat{\mu}_i^{\,2}$
   - RSS is the residual sum of squares and measures the sample variation in the residuals ($\hat{\mu}_i$); that is, how much of the variation is not explained by the model.

The total variation in y can thus always be expressed as the sum of the explained variation and the unexplained variation: $TSS = ESS + RSS$.

GOODNESS-OF-FIT

- Researchers may ask themselves how well does the explanatory or independent variable, X , explain the dependent variable, Y. Or, in other words, how well does the OLS regression line fit the data?
- The R-squared of the regression, also called the coefficient of determination, does just that and is defines as follows:
  - $R^2 \equiv ESS/TSS = 1 - RSS/TSS$
- R^2 is the ratio of the explained variation compared to the total variation in the sample; thus, it is interpreted as the fraction of the sample variation in Y that is explained by X.
- Note that the value of R^2 is always between zero and one, because ESS can be no greater than TSS.

```
> # Vignette 4.2: ----
>
> # What's inside a regression? An intercept, a slope... a line!!
> model <- lm(wage~exper,data=Bwages)
> summary(model)

Call:
lm(formula = wage ~ exper, data = Bwages)

Residuals:
    Min      1Q  Median      3Q     Max
-12.803  -2.554  -0.749   1.643  35.075

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.73486    0.21723   40.21   <2e-16 ***
exper        0.13450    0.01087   12.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.237 on 1470 degrees of freedom
Multiple R-squared:  0.0944,	Adjusted R-squared:  0.09379
F-statistic: 153.2 on 1 and 1470 DF,  p-value: < 2.2e-16
```

The R^2 for that regression equation is 0.094. This means that 9.4% of the variation in Wage in the sample is explained by Experience.

But be careful not to give too much emphasis to the R^2. A low R^2 does not necessarily indicate that a model is “useless.” It may only mean that the phenomenon at hand is just hard to explain.

THE SRM AND CAUSALITY

- Now, does the SRM allow us to draw *ceteris paribus* conclusions about how X affects Y?
- The answer depends on how the unobserved U term relates to the explanatory variable X.
- But the short answer for the SRM is **not likely at all.** Most phenomena are explained by more than one factor, many of which are correlated with the explanatory variable X . (More on this later)

IN-CLASS EXERCISES

Using the ’fertil2’ dataset from ’wooldridge’ on women living in the Republic of Botswana in 1988,

- produce a scatterplot with number of children (children) on the y-axis and education (educ) on the x-axis;
- how do the two variables appear to be related?;
- estimate the regression equation of the number of children on education (note: we say to regress y on x);
- interpret β_0 and β_1;
- plot the regression line on the scatterplot;
- calculate TSS, ESS and RSS. Verify that $TSS = ESS + RSS$;
- using TSS, ESS and RSS, calculate the R^2 of the regression. Verify it is the same as the R^2 reported in the summary of your regression output;
- interpret the R^2 of the regression.