PS2030
Political Research and Analysis
Unit 1: Fundamentals of Linear Regression
1. Bivariate Regression

Spring 2025

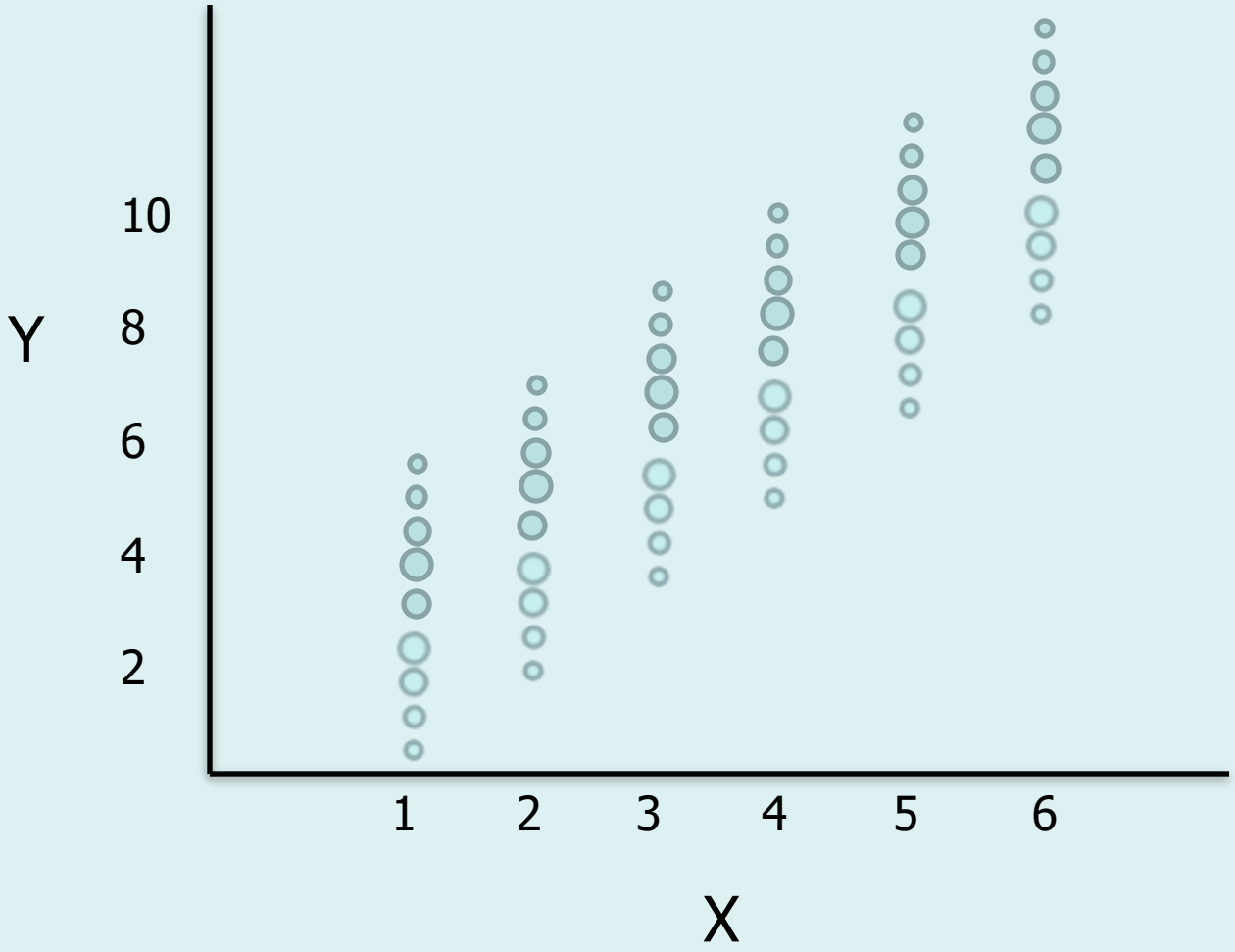
WW Posvar Hall 3600

Professor Steven Finkel

Topics to be Covered

- The population regression function (PRF)
- Fitting a line with sample data (SRF)
- Properties of the ordinary least squares regression line
- Interpretation of OLS coefficients
- Goodness of fit measures

The Relationship Between X and Y in the (Usually Unobserved) Population



X: Years of Education post-High School
Y: Income in $1000s per Month

- How can we characterize this population relationship?
- If we assume a *linear* relationship between X and Y in the population, we can represent the relationship in equation form as:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

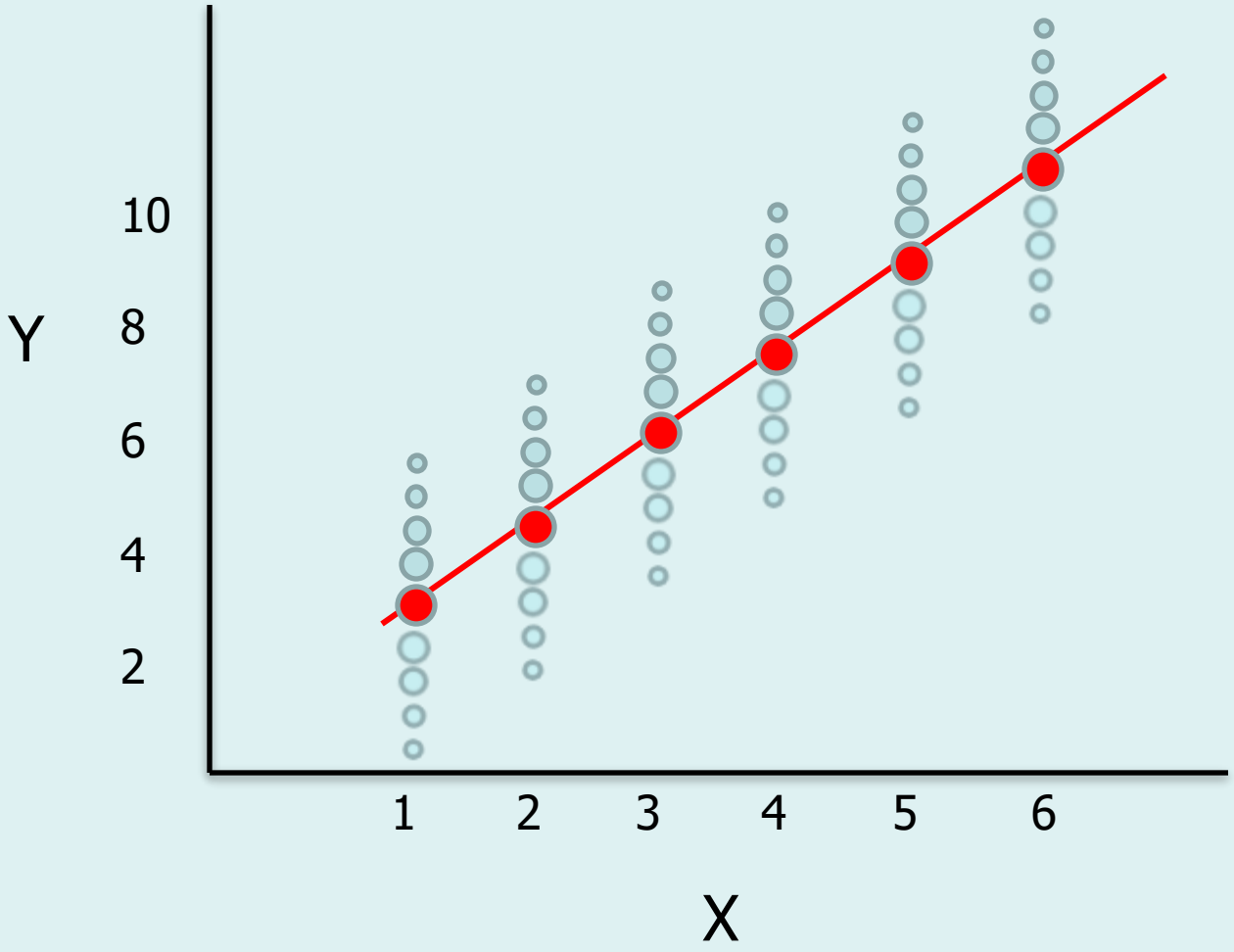
where β_0 is the "intercept" or "starting point" of the line when X=0;

β_1 is the "slope" of the line, or the "average change in Y for every unit change in X"; and

ε_i is the unobserved "error" or "disturbance" term, that portion of Y that is "unexplained" by X

- The line associated with this equation can be seen on the next slide. We call this the "Population Regression Function", in this case the population regression *line*

The Population Regression Function (PRF)



- Formally, the PRF is the function (in this case, a line) that connects the *conditional means* of Y at each level of X *in the population*. (These are the bigger bright red dots on the previous slide).
- The *conditional mean* of Y at a given level of X is represented as:

  $E(Y_i | X)$, or the "Expectation of Y, given X". (An "expectation" is simply a long-run average of some variable or quantity.)
- So:

$$E(Y_i \mid X) = \beta_0 + \beta_1 X_i$$

"For every unit change in X, we change the *conditional mean* of Y by β_1 units". We could also say that "changes in X affect the conditional probability distribution or density function of Y in a systematic, linear fashion".

- We can use this framework to show the "systematic" and "random" parts of Y_i:

$$E(Y_i \mid X) = \beta_0 + \beta_1 X_i, \text{ or}$$

$$\hat{Y}_i = \beta_0 + \beta_1 X_i \qquad \text{Systematic or Deterministic Portion of } Y_i$$

- And

$$Y_i - E(Y_i \mid X) = \varepsilon_i, \text{ or}$$

$$Y_i - \hat{Y}_i = \varepsilon_i, \qquad \text{Random or Stochastic Portion of } Y_i$$

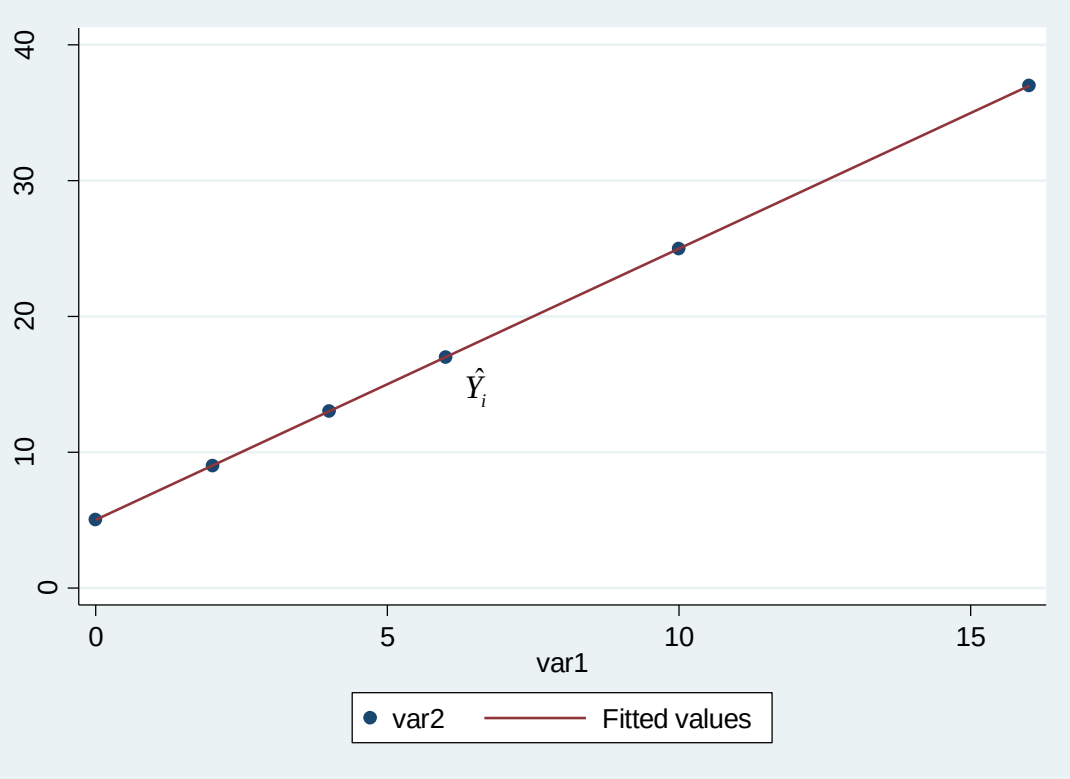
- ε_i is also called the "disturbance" from the population regression function (line), the "unexplained" part of Y
- So we say that Y_i *in the population* in generated through a systematic effect of X ($\beta_0 + \beta_1 X_i$), and a random or stochastic component (ε_i), i.e., a random draw from the distribution of ε_i

What is the Error Term?

If we had a perfectly deterministic relationship, there would be no ε_i at all – all of the observations would fall precisely *on* the population regression function

For each level of X, the conditional mean of Y_i would equal the value of each and every Y_i:

$E(Y_i | X_i) = \beta + \beta X = \hat{Y}_i = Y_i$



- But no social science relationship is deterministic! We will never predict Y perfectly through knowledge of X, i.e., there will always be an ε in our population regression functions. We have *probabalistic*, not deterministic, relationships
- Important to be aware of what ε could potentially represent:
  - **Omitted variables**. All other factors that could have influence on Y. Social science is not so advanced as to think that we have thought of every possible explanatory variable and included it in our predictive models
  - **Random, idiosyncratic influences** on Y (these could be thought of as "variables", but more time or unit specific, conceptually less interesting)
  - **Measurement error**
  - **Human indeterminacy** (and free will?)

Estimating Regression Parameters with Sample Data

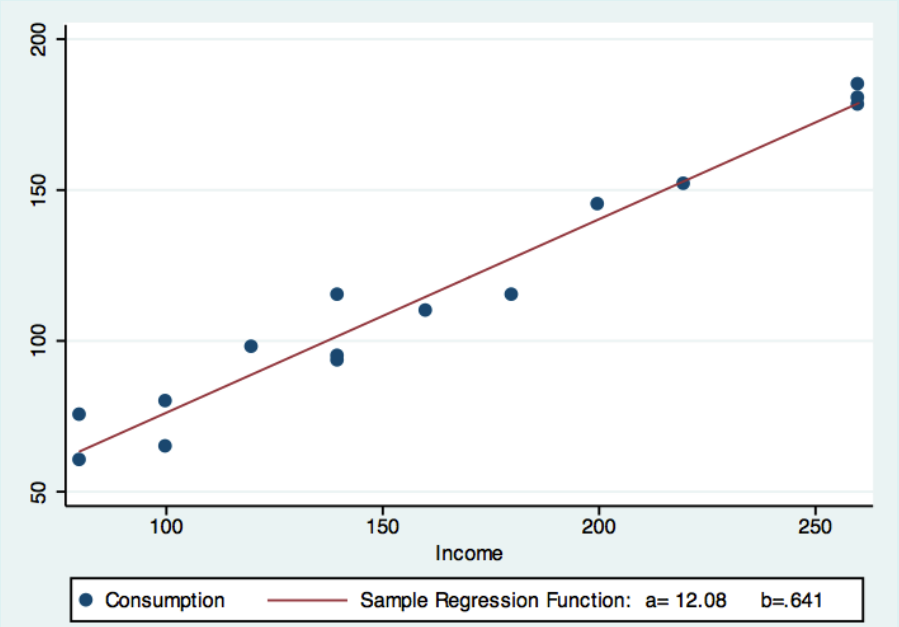
- We don't know the true parameters, i.e. we don't know the values of β_0 and β_1 in the population. We want to estimate these parameters from a sample of data
- We draw a random sample from the population, and plot the graph or "scatterplot" of Y against X
- We can then estimate coefficients in the *sample regression function* (SRF):

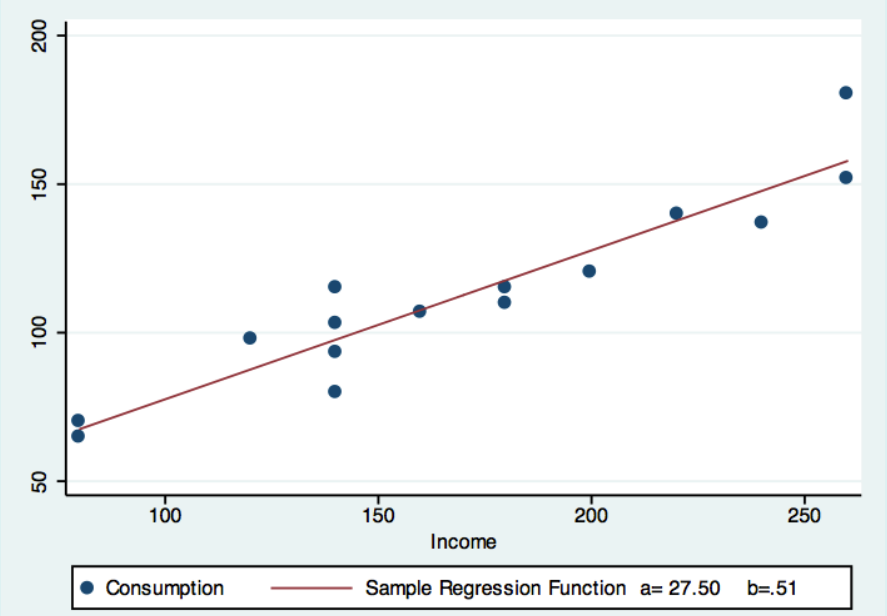
$$\hat{Y}_i = a + bX_i$$

and including a stochastic component yields:

$$Y_i = a + bX_i + e_i$$

- So the goal is to estimate a and b in our sample, and use them to make inferences about population parameters β_0 and β_1
- We therefore want the method we use to arrive at the SRF to get us as close as possible to the (unknown) PRF

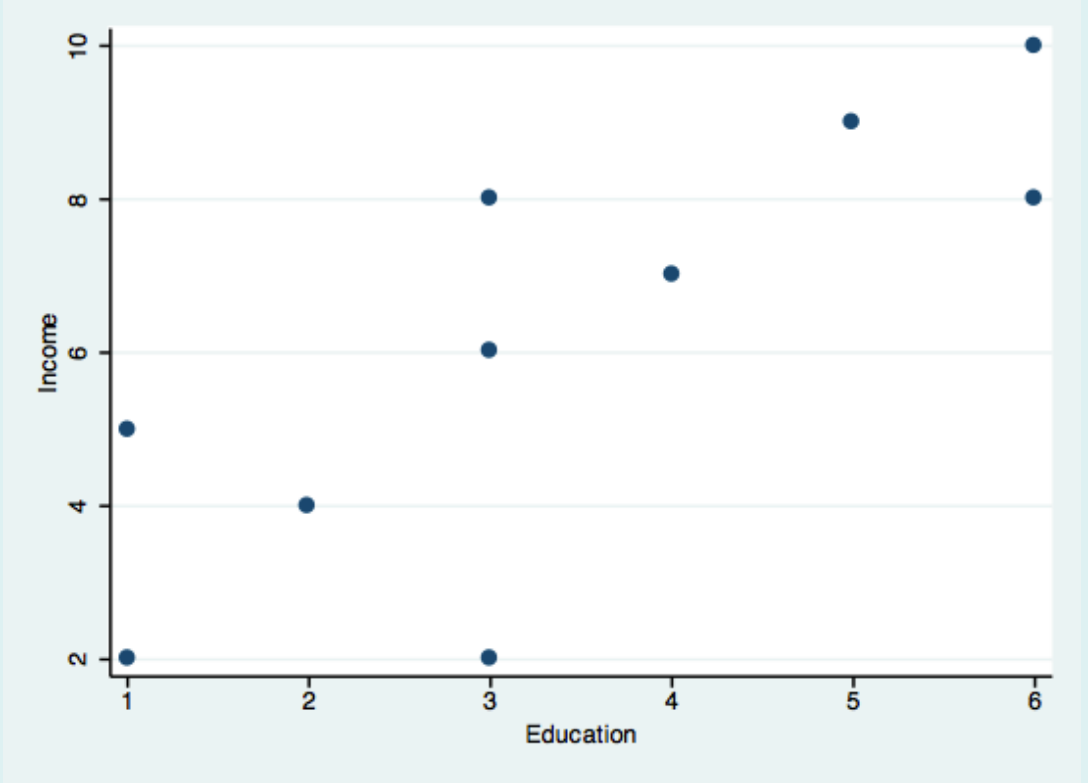




- Important: the coefficients in the SRF will fluctuate because of random sampling from sample to sample. That is, we will get a different estimate of the slope and intercept depending on the particular observations we draw randomly from the population. So we want the method used to estimate the coefficients in the SRF to have some **known properties that will allow us to make accurate inferences about the population parameters**, taking into account random sampling errors
- Let's keep this in mind but defer a more detailed discussion of this point until later!

Estimating the Sample Regression Function

- There are an infinite number of lines that *could* characterize a relationship in a sample
- We need to find some criteria that will allow us to estimate the "best" line that we can
- What do we mean by "best"?
  - The line should come as close as possible to the points, that is, it should somehow *minimize the residuals* from the regression prediction ($\widehat{Y}_i$), i.e., (Y_i -$\widehat{Y}_i$) should be as small as possible
  - The line should have statistical properties permitting the most accurate inferences about the population parameters β_0 and β_1 to be made

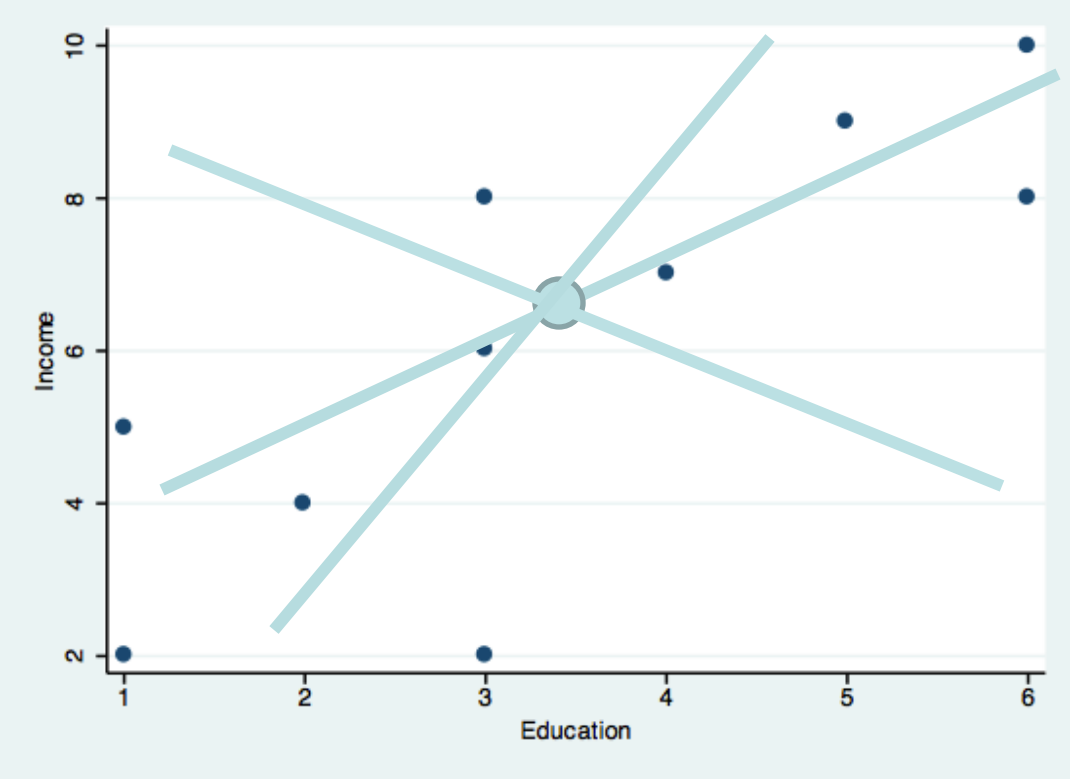


See "ps2030.week1.dta"
Education: Years past High School
Income: In $1000/month

Minimizing the Residuals ($Y_i - \hat{Y}_i$): Alternative Methods

#1: Minimize the Sum of the Residuals, $\Sigma(Y_i - \hat{Y}_i)$

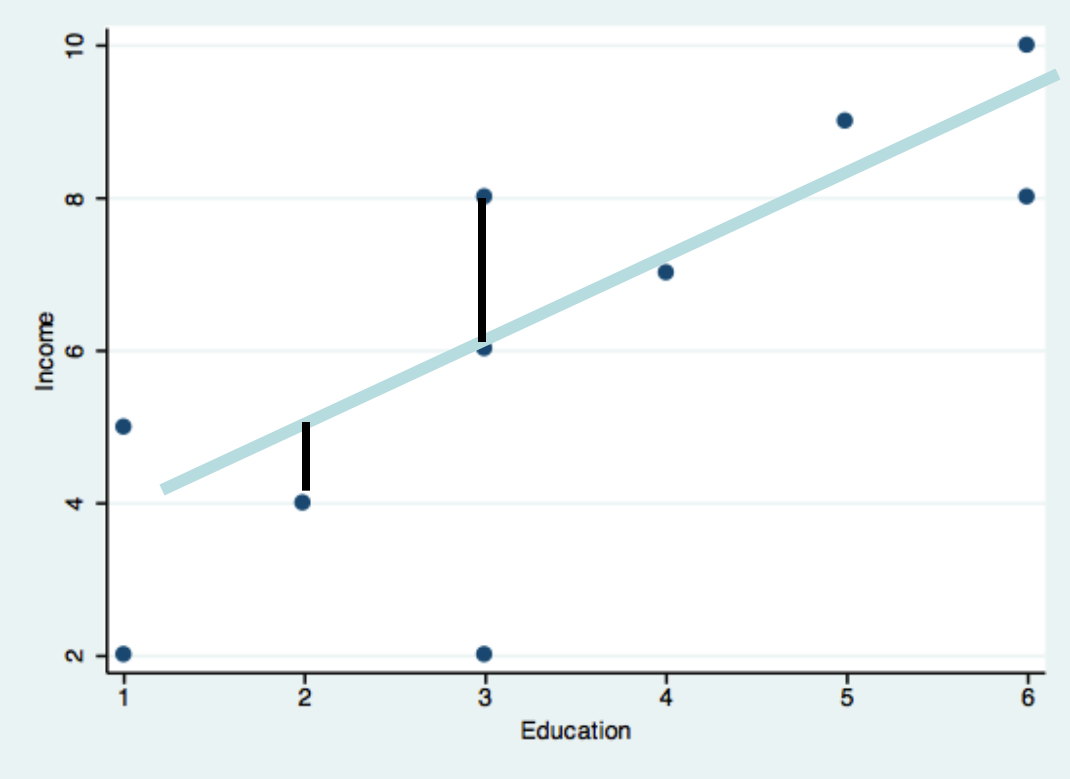
Problem: Any line passing through the points (X-bar, Y-bar) will satisfy this criterion. Negative and positive residuals will cancel out, and the sum of residuals from any of the these (infinite number of) lines will be 0!



Minimizing the Residuals $(Y_i - \widehat{Y}_i)$: Alternative Methods

#2: Minimize the Sum of the Absolute Values of the Residuals, $\Sigma | (Y_i - \widehat{Y}_i) |$

Problem: Negative and positive residuals will not cancel, but this method does not satisfy "best" criteria #2 above, i.e., generating attractive statistical properties for inferences to populations!

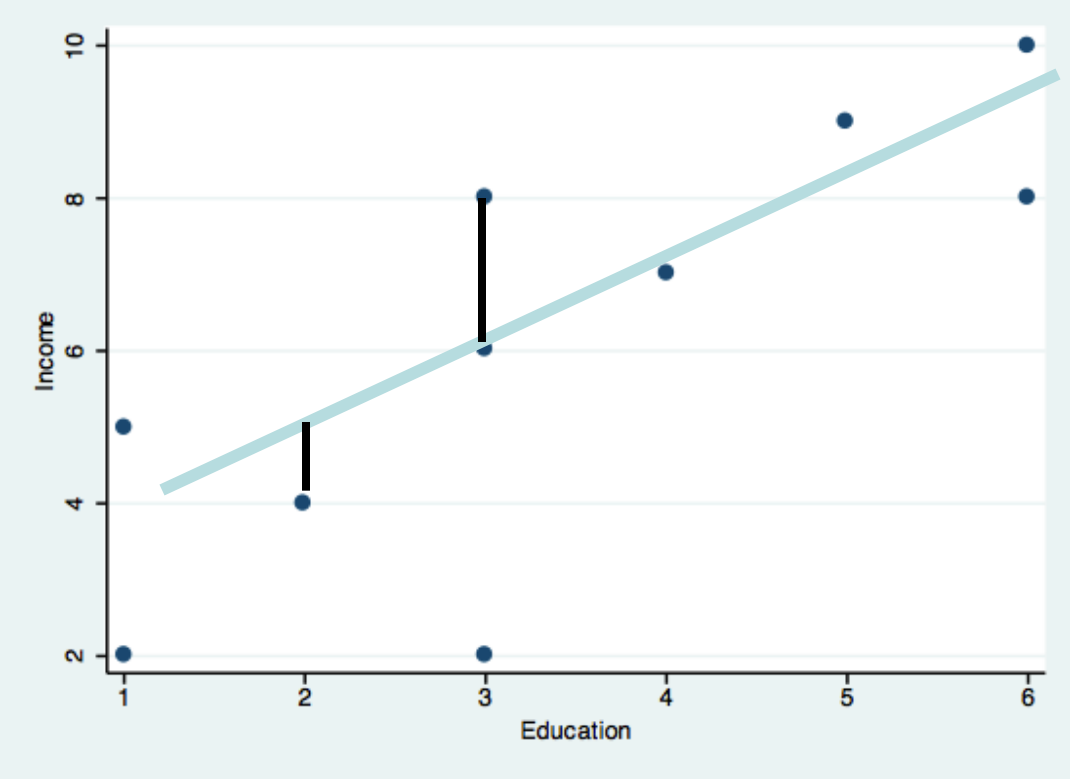


Minimizing the Residuals $(Y_i - \widehat{Y}_i)$: Alternative Methods

#3: Minimize the Sum of the Squared Values of the Residuals, $\Sigma(Y_i - \widehat{Y}_i)^2$

Negative and positive residuals do not cancel, and this method has many attractive statistical properties for inferences to populations!

This estimation method for the SRF is called **"Least Squares"**– it estimates the line which produces the smallest "sums of squares of residuals", or the fewest total squared deviations between the actual Y values and the predicted Y values from the line



Estimating the SRF (intercept "a" and slope "b") via Least Squares

- Goal: minimize $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, or $\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$
- Which means minimizing $\sum_{i=1}^{n}(a + bX_i - Y_i)^2$
- A quadratic function is minimized when the partial derivatives, in this case, with respect to "a" and "b", are set to zero:

$$wrt\ a:\ 2\Sigma(a + bX_i - Y_i) = 0$$

$$wrt\ \text{b}:\ 2\Sigma X_i(a + bX_i - Y_i) = 0$$

- Remember all this from previous class(es)?"

Summation Algebra

- To solve the equations for "a" and "b", we apply the rules of summation algebra (which will come in handy in other areas of the course as well)

- 1) $\sum_{i=1}^{n}(a) = Na$ "the sum of a constant is N times the constant"

- 2) $\sum_{i=1}^{n}(ax) = a\sum_{1}^{n}x$ "the sum of a constant times a variable is the constant times the sum of the variable"

- 3) $\sum_{i=1}^{n}(a+x) = Na + \sum_{1}^{n}x$ "the sum of a constant plus a variable is N times the constant plus the sum of the variable"

- 4) $\sum_{i=1}^{n}(x+y) = \sum_{1}^{n}x + \sum_{1}^{n}y$ "the sum of two variables is equal to the sum of the two separately"

- Note relationship between "summations" and "expectations"

- Carrying out this algebra yields what are called the two "Normal Equations" of least squares regression:

$$(1)\quad Na + b\Sigma X_i = \Sigma Y_i$$

$$(2)\quad a\Sigma X_i + b\Sigma X_i^2 = \Sigma X_i Y_i$$

- Dividing (1) by N gives: $a = \bar{Y} - b\bar{X}$

- and substituting this expression for "a" in (2) gives:

$$b = \frac{\Sigma(X_i Y_i) - \frac{\Sigma X_i \Sigma Y_i}{N}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{N}}$$

These are the "Least Squares" estimates of the slope and intercept!

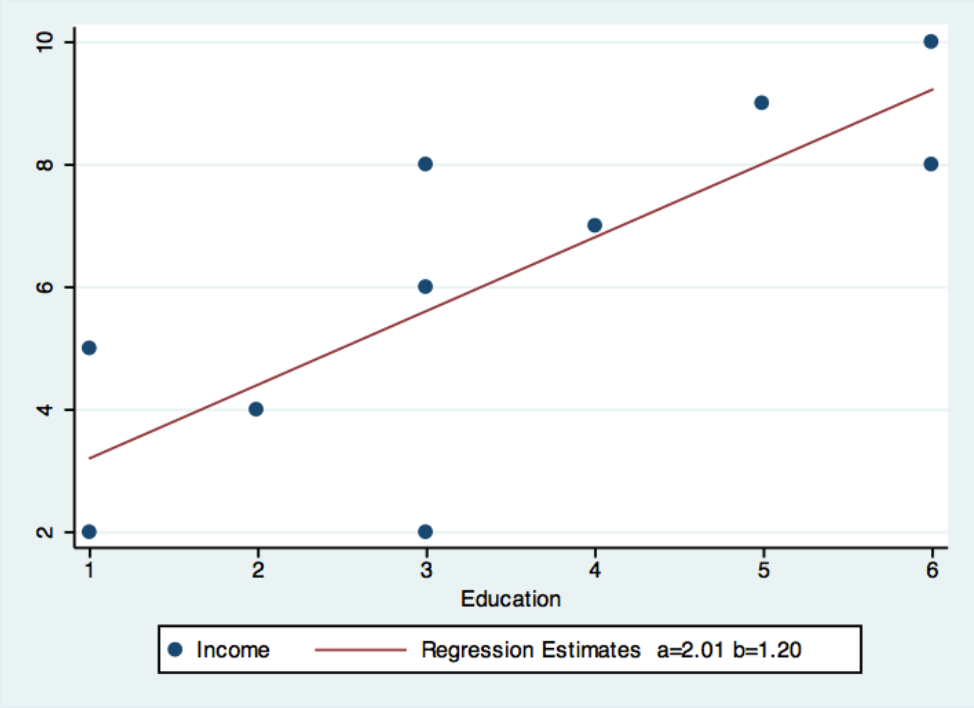
or

$$b = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$$

Our Example

- See Excel file "Example of OLS Regression Calculations.2024.xlsx"

- a=2.01
- b=1.20




```
. regress var2 var1
```

Source	SS	df	MS
Model	44.0644737	1	44.0644737
Residual	26.8355263	8	3.35444079
Total	70.9	9	7.87777778

Number of obs	10
F(1, 8)	13.14
Prob > F	0.0067
R-squared	0.6215
Adj R-squared	0.5742
Root MSE	1.8315

var2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
var1	1.203947	.3321798	3.62	0.007	.4379393	1.969955
_cons	2.006579	1.269257	1.58	0.153	-.9203338	4.933492

Interpretations

- "a" Interpretation: the predicted place on the line for Y when X=0
  - May not be realistic to have X=0 so be cautious in this interpretation. It is a mathematical necessity but substantively not always meaningful
- "b" Interpretation: as X changes by 1 unit, Y changes *on average* by β units. This is the "estimated effect of X on Y"
  - We will not necessarily be accurate in predicting individual values of Y, and one of the measures of the "strength" of the relationship will be how accurate in fact we are in predicting Y from X (see next slides)
  - But b nevertheless represents our best estimate of the *substantive* relationship between X and Y, expressed in the given units of Y. In our example, each additional year of education gives you 1.2 thousands of dollars more in income
  - ALWAYS INTERPRET THIS VALUE SUBSTANTIVELY – IS IT A LOT, A LITTLE, A BIG EFFECT, SMALL, OR WHAT?

Nice Properties of OLS Regression Line

- The sum of squared residuals from the line is at a minimum

  (*by definition*)

- The sum of residuals is 0

$$\Sigma(Y_i - \hat{Y}_i) = \Sigma e_i$$

$$\Sigma(Y_i - a - bX_i) = \Sigma e_i$$

$$\Sigma Y_i - \Sigma a - b\Sigma X_i = \Sigma e_i$$

$$\Sigma Y_i - Na - b\Sigma X_i = \Sigma e_i$$

from first normal equation, $\Sigma Y_i = Na + b\Sigma X_i$

$$Na + b\Sigma X_i - Na - b\Sigma X_i = \Sigma e_i$$

$$0 = \Sigma e_i$$

- Regression line passes through $\bar{X}, \bar{Y}$

$$\hat{Y}_i = a + bX_i$$

$$a = \bar{Y} - b\bar{X}$$

$$\hat{Y}_i = \bar{Y} - b\bar{X} + bX_i$$

$$\hat{Y}_i = \bar{Y} + b(X_i - \bar{X})$$

So when $X_i = \bar{X}$, $\hat{Y}_i = \bar{Y}$

- **This expresses the important idea of "covariation" in regression**

$\hat{Y}_i$ is equal to the mean of Y plus some weighted distance of X_i from its mean (with "b" being the weight). So Y's predicted distance from its mean depends to some (weighted) extent on X's distance from its mean. This is intuitively what regression gives you, the "weight" in this sense.

- Residuals from OLS are uncorrelated with X, i.e., $\Sigma(X_i e_i) = 0$

$$e_i = Y_i - a - bX_i$$

$$\Sigma X_i e_i = \Sigma(X_i(Y_i - a - bX_i))$$

$$\Sigma X_i e_i = \Sigma(X_i Y_i - X_i a - X_i bX_i)$$

$$\Sigma X_i e_i = \Sigma X_i Y_i - a\Sigma X_i - b\Sigma X_i^2$$

from second normal equation, $\Sigma X_i Y_i = a\Sigma X_i + b\Sigma X_i^2$

$$\Sigma X_i e_i = a\Sigma X_i + b\Sigma X_i^2 - a\Sigma X_i - b\Sigma X_i^2$$

$$\Sigma X_i e_i = 0$$

- This means that OLS gives you a line where the errors of prediction are uncorrelated *by construction* with the level of X. This makes sense: if you had a line where low levels of X means the line is overpredicting Y (i.e., low residuals) and high levels of X mean the line is underpredicting Y (i.e. high residuals), you would want the "best" line to adjust for this!! **(Unless X and ε truly are related in the population – this is very important!)**
- Residuals from OLS are uncorrelated with Y as well, i.e. $\Sigma Y_i e_i = 0$

Additional Notes on Bivariate Regression

- Note "b" formula: $\frac{\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2}$; this is equivalent (if divide both the numerator and the denominator by N) to the

  "***Covariance of X and Y, divided by the Variance of X***"

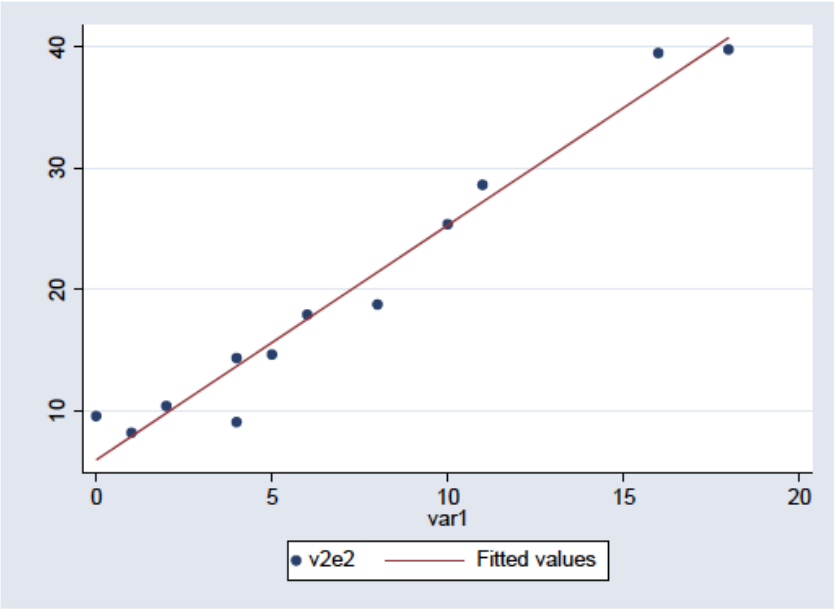
- "b" is our estimate of the causal effect of X on Y, but what we have done so far *almost never* gives us the true picture of the causal relationship

  - Sampling Error – β_1 in the overall population could still be 0
  - Reciprocal causality in non-experimental designs
  - Spuriousness caused by omitted variables X2, X3, X4, etc.!!!!
  - Other kinds of model misspecification, e.g., the relationship may not be linear (should always check the initial scatterplot to give some indication of this possibility)

Assessing the Strength of a Regression Relationship

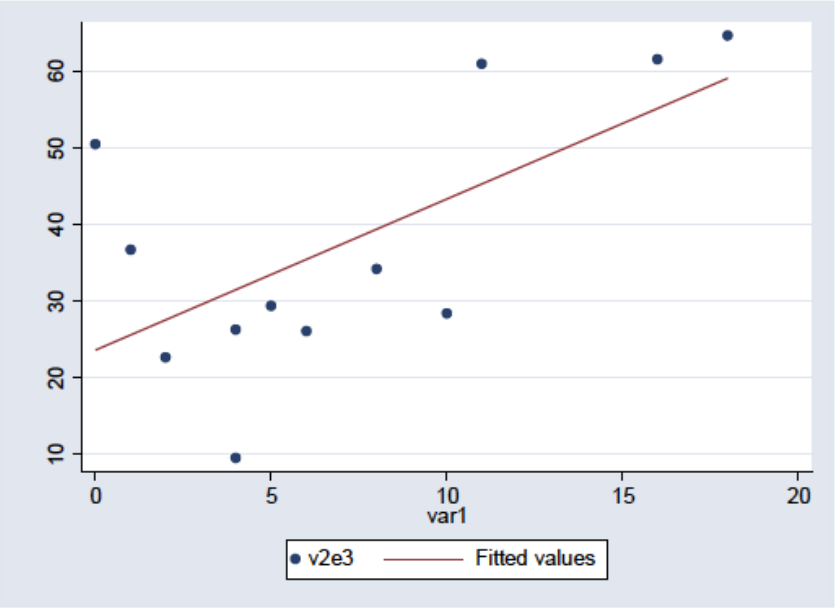
- Basic Question: How close are the points, on average, from the regression line? If very close, we have a "strong" relationship; if not close, we have a "weak" relationship
- Simplest measure of association in this regard is called the *correlation coefficient*, or "Pearson's correlation." Runs from -1, for a perfect negative relationship, to 0 for no relationship, to +1 for a perfect positive relationship
- Formula: $$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

Which, if divide the numerator and denominator both by N, is the "Covariance of X and Y, divided by the product of the Standard Deviation of X and the Standard Deviation of Y'. It is the "standardized covariance" between X and Y


40
30
20
10
0
5
10
15
20
var1
v2e2
Fitted values

← r=.98

r=.64 →


60
50
40
30
20
10
0
5
10
15
20
var1
v2e3
Fitted values

Problems/Issues with Pearson's R

- Gives crude sense of correlation only, no exact "unit change" in Y interpretation like β
- Bidirectional statistic, in that r(XY)= r(YX), so says nothing about causality (reflecting the truism that "correlation does not equal causality")
- We don't know what the meaning of the -1 to 1 scale is, aside from "bigger numbers means stronger positive or negative relationship", and "numbers closer to 0 mean a weak or no relationship". Doesn't tell us as much as measures based on PRE (Proportional Reduction in Error) principle that we will discuss.

- But a useful bivariate relationship: b=r*SD(Y)/SD(X), since

$$b = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \text{ and } \quad r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma(X_i - \bar{X})^2 \Sigma(Y_i - \bar{Y})^2}}$$

More Useful Measure: "R-squared"

- Question: How much of the total variation of Y around its mean is "explained" by X, and how much is residual or "unexplained" variation?

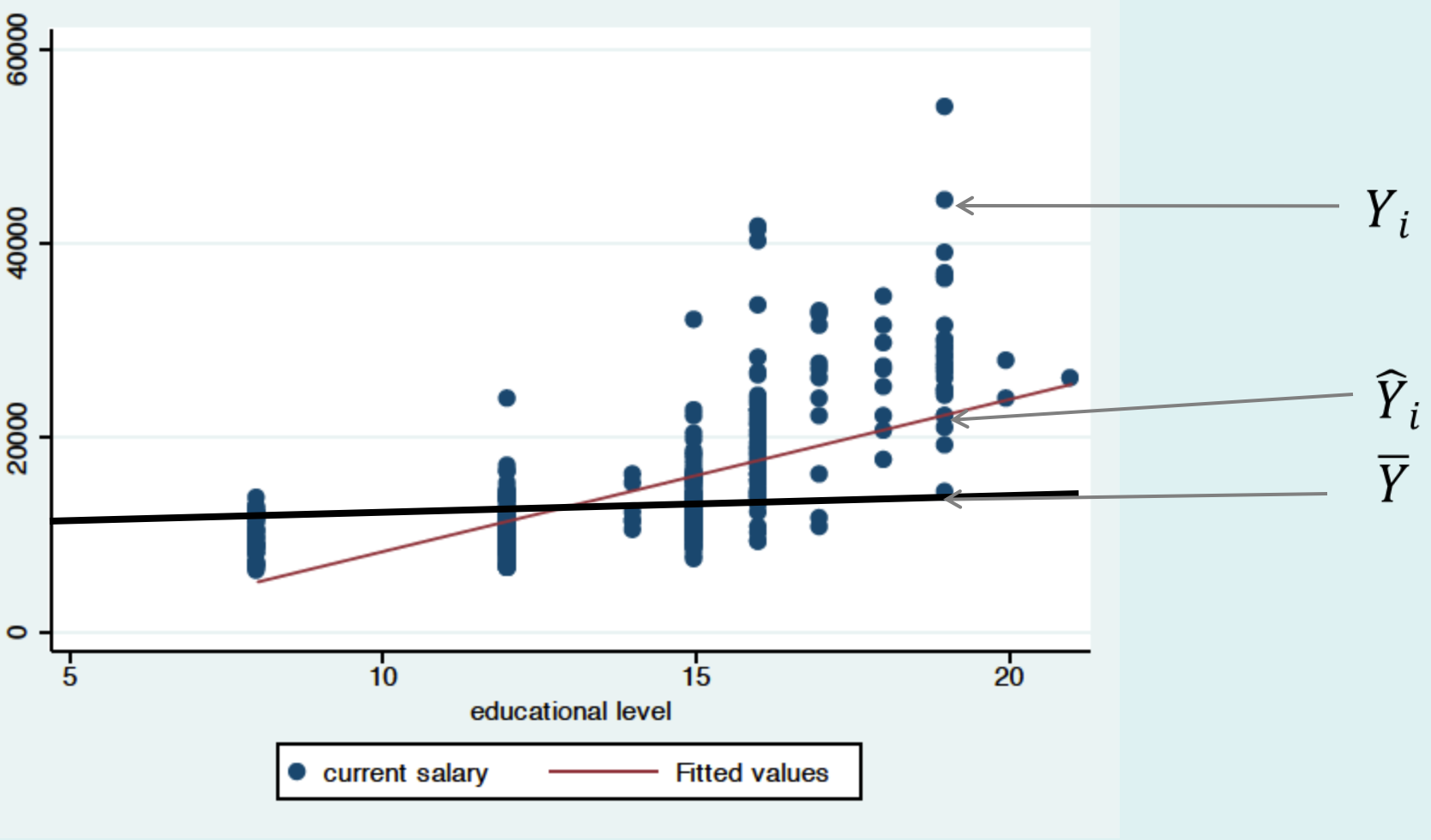
  Total Sum of Mean Deviations in Y: $\sum Y_i - \bar{Y}_i$

  Total Sum of Residuals from OLS Prediction: $\sum Y_i - \hat{Y}_i$

  Total Sum of Deviations from OLS Prediction to Mean of Y: $\sum \hat{Y}_i - \bar{Y}_i$

The total deviation in Y_i can be partitioned into the two parts:

$$\sum(Y_i - \bar{Y}_i) = \sum(Y_i - \hat{Y}_i) \quad - \quad \sum(\hat{Y}_i - \bar{Y}_i)$$


60000
40000
20000
0
5
10
15
20
educational level
current salary
Fitted values
Y_i
$\hat{Y}_i$
$\overline{Y}$

- We can show that:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(\hat{Y}_i - \bar{Y})^2 + \Sigma(Y_i - \hat{Y}_i)^2$$

Total Sum of Squares in Y = Regression Sum of Squares + Residual Sum of Squares

- We calculate "R-squared" as the proportion of "regression" or "explained" sum of squares, or "explained" squared deviations in Y:

$$R^2 = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

- It goes from 0 to 1 and is interpreted exactly as a proportion (i.e. 20% of the total squared deviations in Y are explained by X if R-squared is .2, 40% of the total squared deviations in Y are explained by X if R-squared is .4, etc.)

"PRE" Interpretation of R-squared

- R-squared also has a nice interpretation in terms of the "proportional reduction in errors" idea, i.e., how much we can reduce our predictive errors in Y through knowledge of X.
- Many measures of association for nominal or ordinal variables with which you may be familiar are PRE measures, e.g., lambda for nominal variables and gamma for ordinal variables.
- Basic calculation of all PRE measures:

  (Errors in Predicting Y Without Knowledge of X -

  Errors in Predicting Y with Knowledge of X)/

  Errors in Predicting Y without Knowledge of X

  or (Errors Without X-Errors With X)/Errors Without X)
- With R-squared, we talk about "*squared errors* of prediction in Y"

- (Squared) Errors of Prediction in Y without Knowledge of X:

  $\Sigma(Y_i - \bar{Y})^2$ (we would guess the mean of Y for all Y_i)
- (Squared) Errors of Prediction in Y with Knowledge of X:

  $\Sigma(Y_i - \hat{Y}_i)^2$ (we would guess the predicted Y from the regression line for all Y_i, and the residual would be our error in prediction)
- So: $$R^2 = \frac{\Sigma(Y_i - \bar{Y})^2 - \Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2} \quad \text{or} \quad 1 - \frac{\Sigma(Y_i - \hat{Y}_i)^2}{\Sigma(Y_i - \bar{Y})^2} \quad \text{or} \quad \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2}$$

- "Through knowledge of X (the regression line), we have reduced our squared errors of prediction in Y by" a proportion of whatever value R-squared is
- Since we use OLS, the squared errors of prediction with X is at a minimum, which means that OLS maximizes R-squared!

Notes on R-squared

- $R^2 = (r)^2$; that is, R^2 **for a bivariate regression** is equal to Pearson r, squared
- A low R^2 either means that you have a poor explanatory model, or the relationship is possibly non-linear, or you have little variation in your DV to explain in the first place
- Some research is less concerned with model fit than with precise estimates of causal effect of one variable on another (i.e., rigorous causal inference and explanatory power of the model are different issues)
- A high R^2 either means that you have a good model, or that you have a good statistical model but not necessarily a meaningful substantive model (e.g. Y at time t predicted by Y at time t-1)

- **Therefore, R^2 is a highly informative and useful measure of overall model fit, but its importance should not be overestimated**

Final Measure of "Strength": The Root Mean Square Error (RMSE), or the Standard Error of Estimate (SEE)

- Useful to have an estimate of the association between X and Y that is expressed in terms of the units of Y – that is, how far off are we on average in predicting Y with X?
  - For example: in predicting a person's political participation, are we off by 5 behaviors? 1? .5? Obviously the smaller our predictive error on average, the stronger our relationship
  - Analogous logic to R-squared, but this measure would be expressed in the units of the dependent variable

- "**Standard Error of Estimate**", or "**Root Mean Squared Error**", is calculated as:

$$RMSE = \sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}}$$

- RMSE is the square root of the average squared error of prediction, or the square root of the error variance

  Error: $(Y_i - \hat{Y}_i)$

  Error Sums of Squares: $\Sigma(Y_i - \hat{Y}_i)^2$

  Error Variance: $\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2} = \hat{\sigma}^2$

  RMSE=$\sqrt{\frac{\Sigma(Y_i - \hat{Y}_i)^2}{N-2}} = \hat{\sigma}$

- Or, RMSE is the "standard deviation of the residuals"

- Compare to the overall standard deviation of Y (SD(Y) to see how much X has caused a reduction in standard deviation of Y, *conditioned on X* (SD(Y | X)

- This gives a goodness of fit measure in the units of the dependent variable: we are off on average in predicting Y by whatever the RMSE or SEE is
- So, we can now answer several questions about the relationship between X and Y:
  - What is the "effect" of X on Y in terms of the average change in Y produced by a unit change in X (β_1)?
  - What is the strength of the association between X and Y, that is, do they covary together and can we reduce our errors in prediction in Y through knowledge of X (r, R-squared, and RMSE)?
  - How far off on average are we in predicting Y with X (RMSE)?

- Note differences in all these measures, but there are some close mathematical relations (and some conceptual relations too, given that r, R-squared and β are all rooted in the idea of covariation between X and Y – see earlier slide under "Properties of OLS line")

$$b = \frac{r * SD(Y)}{SD(X)}$$

$$r = \frac{b * SD(X)}{SD(Y)}$$

$$R^2 = \frac{b^2 Var(X)}{Var(Y)} = \frac{b^2 \mathsf{S}(X_i - \bar{X})^2}{\mathsf{S}(Y_i - \bar{Y})^2} = \frac{\mathsf{S}(\hat{Y}_i - \bar{Y})^2}{\mathsf{S}(Y_i - \bar{Y})^2} = r(Y_i, \hat{Y}_i)^2$$