

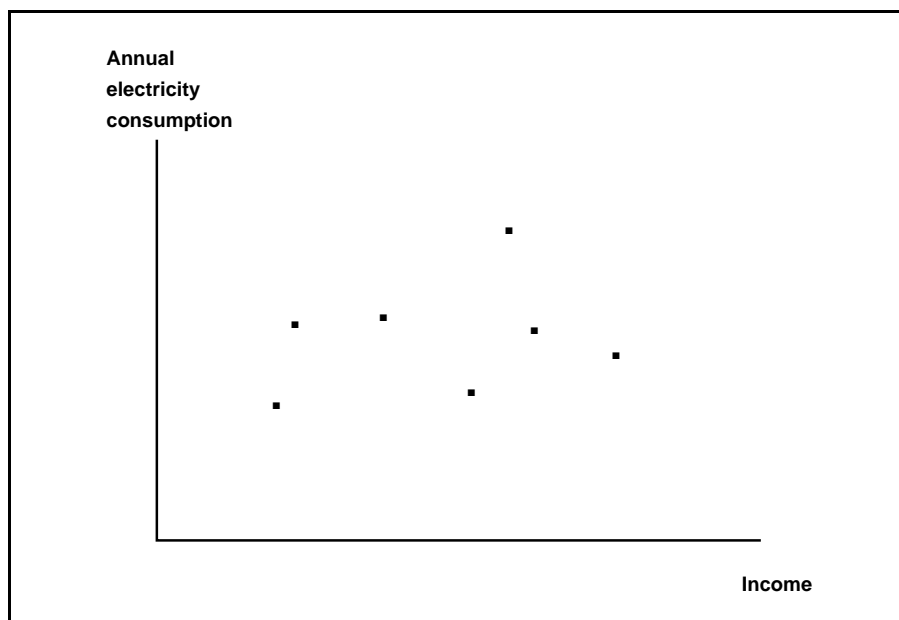
LECTURE / DISCUSSION

Regression

Regression as Data Fitting

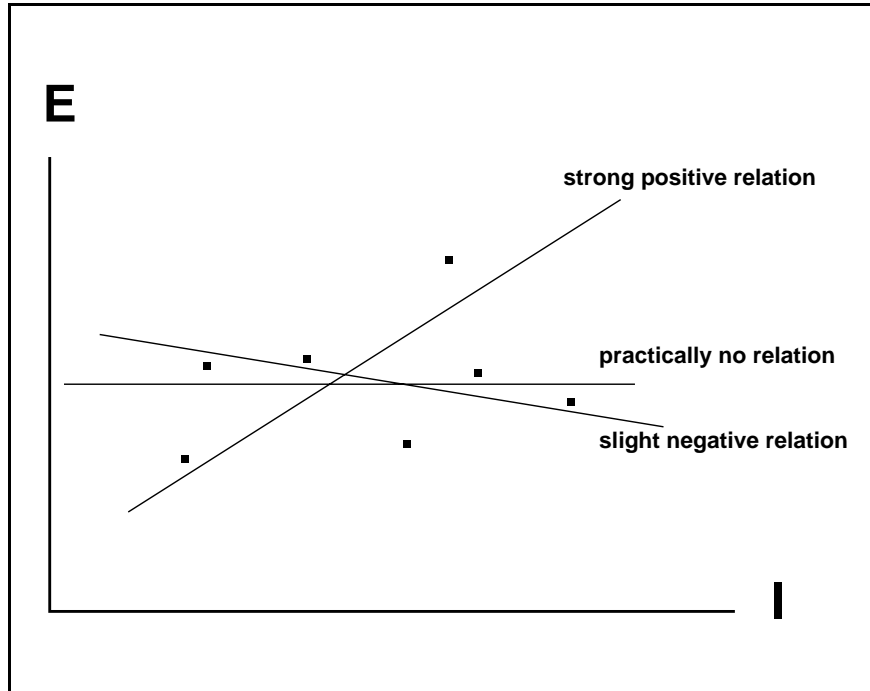
Question: Do high income households consume more or less electricity than lower income households?

To answer: Take a sample of households. Observe the energy consumption and income of each household.

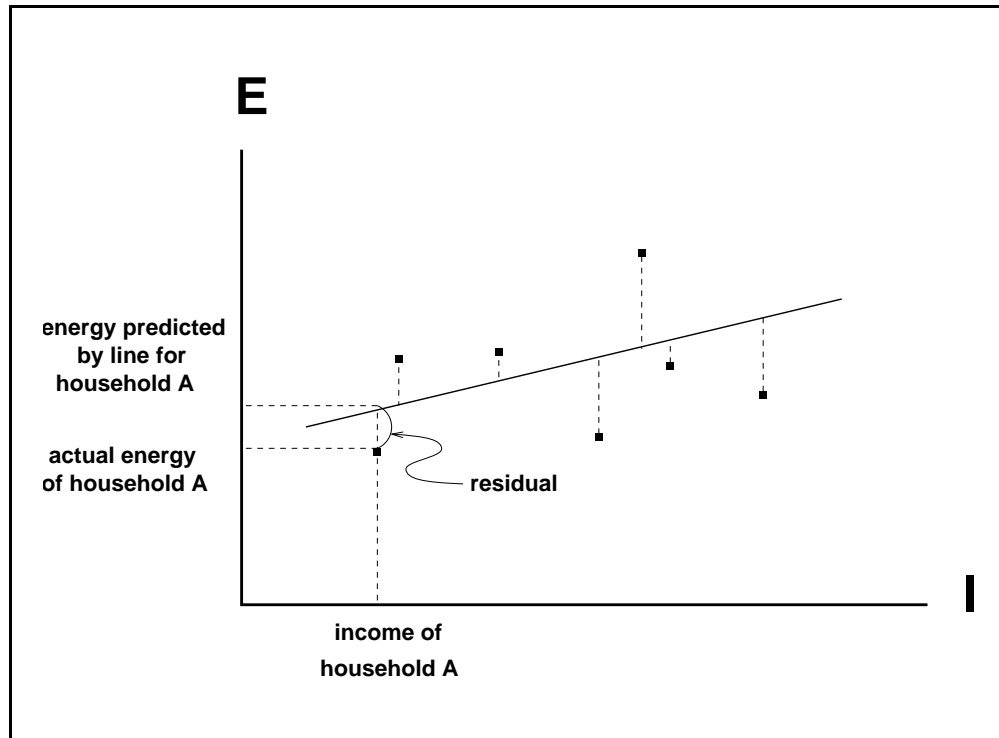


- We want a line that summarizes the general trend of the data.
- What is the best line? What line fits the data best?

Which line is best?



Definition of Residuals

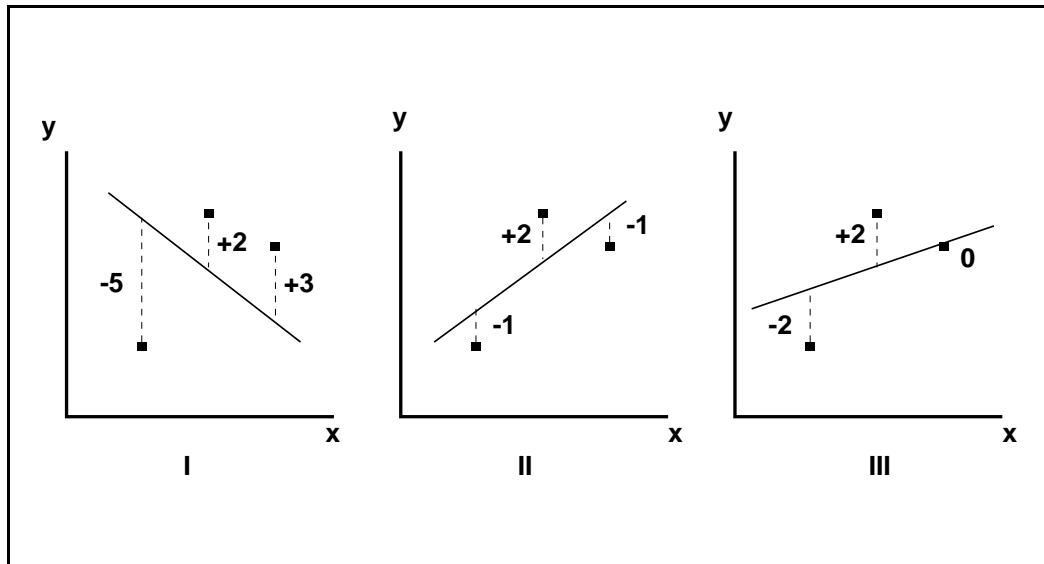


Line: $\hat{E}_n = \alpha + \beta I_n$; I denotes income and
n denotes household;
 $n = 1, \dots, N$

Residual: $r_n = E_n - \hat{E}_n$

For a good fit, we want residuals to be small.

Which residuals are smallest?



Candidate criteria:

1. The best line gives the smallest *sum of residuals*.

$$\min \left| \sum_{n=1}^N (E_n - \hat{E}_n) \right|$$

Problem: positives and negatives cancel out.

$$\text{I. } -5 + 2 + 3 = 0$$

$$\text{II. } -1 + 2 - 1 = 0$$

$$\text{III. } -2 + 2 + 0 = 0$$

All lines are equally good by this criterion.

2. The best line gives the smallest ***sum of absolute deviations***

$$\min \sum |E_n - \hat{E}_n|$$

Problem: cannot always distinguish lines.

I. $5 + 2 + 3 = 10$

II. $1 + 2 + 1 = 4$

III. $2 + 2 + 0 = 4$

Also, difficult mathematically.

3. The best line gives the smallest ***sum of squared residuals***.

$$\min \sum (E_n - \hat{E}_n)^2$$

I. $25 + 4 + 9 = 38$

II. $1 + 4 + 1 = 6$

III. $4 + 4 + 0 = 8$

- OLS:
- weights outliers more
 - easy mathematically
 - is unbiased and efficient (shown later)

Derivation of OLS Line

Observations: X_n, Y_n for $n = 1, \dots, N$

Line: $\hat{Y}_n = \alpha + \beta X_n$

$$\text{SSR: } \sum_{n=1}^N (Y_n - \hat{Y}_n)^2 = \sum_{n=1}^N (Y_n - \alpha - \beta X_n)^2$$

Find α and β that minimize SSR.

Find OLS $\hat{\alpha}$

$$SSR = \sum (Y_n - \alpha - \beta X_n)^2$$

$$\frac{\partial SSR}{\partial \alpha} = - \sum 2(Y_n - \alpha - \beta X_n) = 0$$

$$- \sum Y_n + N\alpha + \beta \sum X_n = 0$$

$$N\alpha = \sum Y_n - \beta \sum X_n$$

$$\hat{\alpha} = \bar{Y} - \beta \bar{X}$$

where bar denotes sample mean.

Find OLS $\hat{\beta}$

Given $\hat{\alpha}$, we can rewrite residuals

$$\begin{aligned}r_n &= Y_n - \hat{\alpha} - \beta X_n = Y_n - (\bar{Y} - \beta \bar{X}) - \beta X_n \\ &= (Y_n - \bar{Y}) - \beta(X_n - \bar{X}) \\ &= y_n - \beta x_n\end{aligned}$$

where lower case letters denote deviations from the sample means.

$$SSR = \sum (y_n - \beta x_n)^2$$

$$\frac{\partial SSR}{\partial \beta} = - \sum 2(y_n - \hat{\beta} x_n) x_n = 0$$

$$- \sum y_n x_n + \hat{\beta} \sum x_n^2 = 0$$

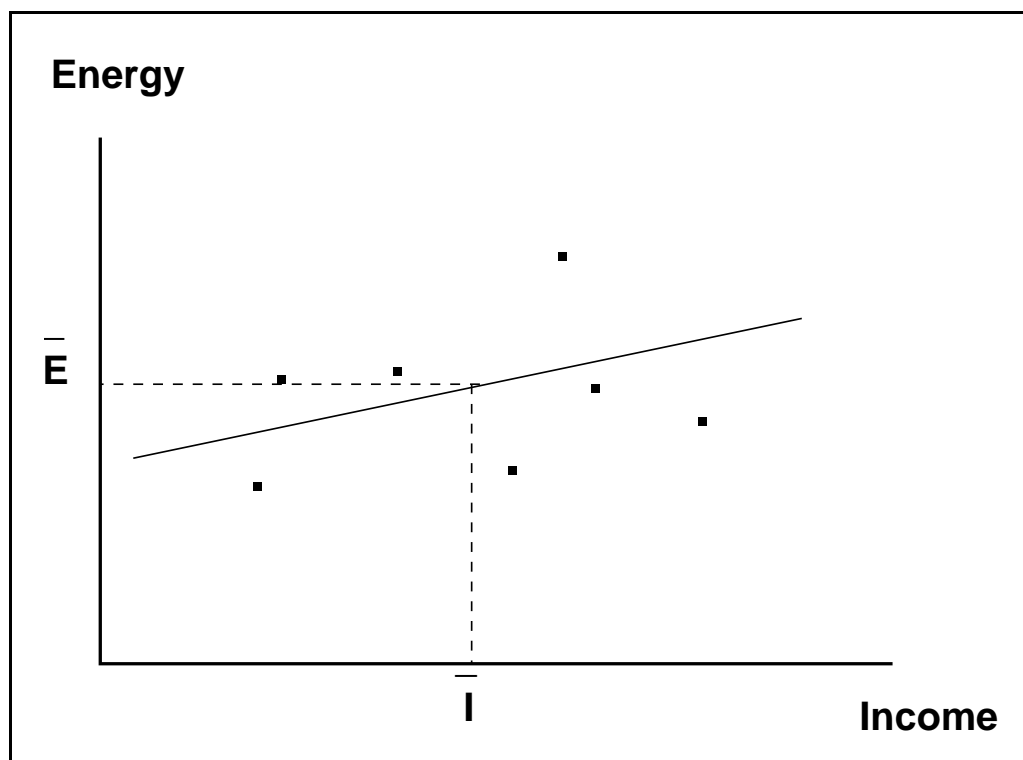
$$\hat{\beta} \sum x_n^2 = \sum y_n x_n$$

$$\hat{\beta} = \frac{\sum y_n x_n}{\sum x_n^2}$$

$$\hat{\beta} = \frac{\text{cov}(y, x)}{\text{var}(x)}$$

Properties of OLS Line

Property 1. OLS line goes through midpoint of data



$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

Evaluate \hat{Y} at average \bar{X} :

$$\hat{\alpha} + \hat{\beta}\bar{X} = (\bar{Y} - \hat{\beta}\bar{X}) + \hat{\beta}\bar{X} = \bar{Y}$$

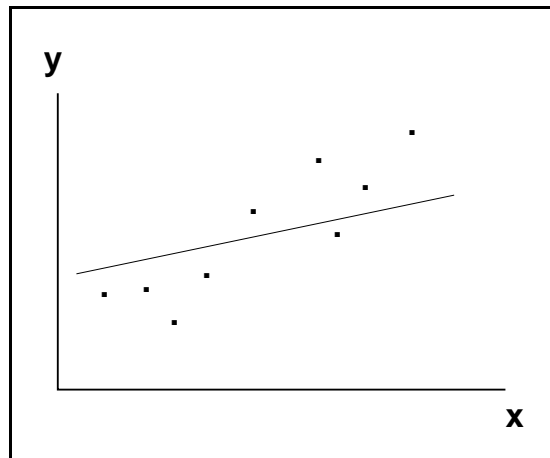
Property 2. The average fitted residual is zero.

$$\begin{aligned}\bar{r} &= \frac{1}{N} \sum r_n = \frac{1}{N} \sum (Y_n - \hat{Y}_n) \\ &= \frac{1}{N} \sum Y_n - \frac{1}{N} \sum (\hat{\alpha} + \hat{\beta} X_n) \\ &= \bar{Y} - (\hat{\alpha} + \hat{\beta} \bar{X}) \\ &= \bar{Y} - \bar{Y} = 0\end{aligned}$$

☞ Line is accurate on average.

Property 3. Fitted residuals are uncorrelated with explanatory variables.

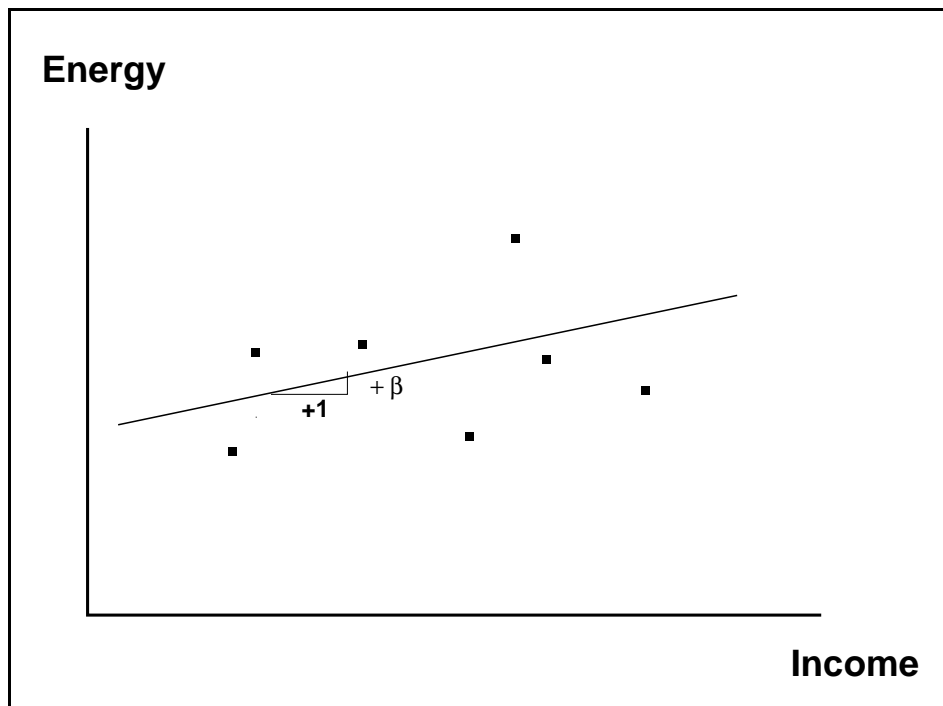
Suppose they are correlated.



A better fit can be obtained by changing the slope.

$$\begin{aligned}\text{cov}(r,x) &= \frac{1}{N} \sum r_n x_n = \frac{1}{N} \sum (y_n - \hat{\beta} x_n) x_n \\ &= \frac{1}{N} \sum y_n x_n - \hat{\beta} \frac{1}{N} \sum x_n^2 \\ &= \text{cov}(x,y) - \hat{\beta} \text{var}(x) \\ &= \text{cov}(x,y) - \frac{\text{cov}(x,y)}{\text{var}(x)} \cdot \text{var}(x) \\ &= 0\end{aligned}$$

Property 4. Slope of OLS line gives how much y is predicted to change when x changes by one unit. So: $\hat{\beta}$ depends on units of x and y .



Example: $\hat{E} = \hat{\alpha} + 9 \times I$

\$1 extra income is predicted by line to induce household to buy 9 more kWhs.

Implication of property 4: If units of x or y are changed, $\hat{\beta}$ will adjust appropriately.

Example:

Measure income in **thousands** of dollars instead of dollars:

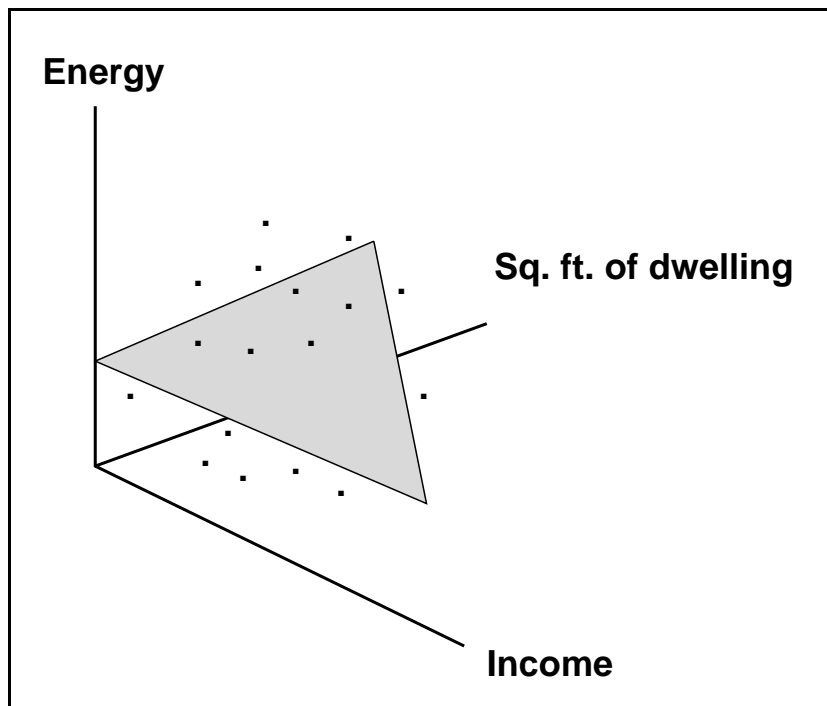
$$\$10,000 \Rightarrow I = 10$$

OLS line becomes:

$$\hat{E} = \alpha + 9000I$$

$$\begin{aligned}\hat{\beta} &= \frac{\sum y_n x_n}{\sum x_n^2} = \frac{\sum y_n (x_n/1000)}{\sum (x_n/1000)^2} \\ &= \frac{\sum y_n x_n (1/1000)}{\sum x_n^2 (1/1000)^2} \\ &= \frac{\sum y_n x_n}{\sum x_n^2} \cdot \frac{1}{1/1000} \\ &= 1000 \frac{\sum y_n x_n}{\sum x_n^2}\end{aligned}$$

Two Explanatory Variables



$$\hat{E} = \alpha + \beta I + \theta S$$

OLS fits a plane that minimizes SSR.

$\hat{\beta}$ gives effect of income holding dwelling size constant.

How Good is the Fit?

R^2 gives percent of variation in dependent variable that is explained by the model.

$$\text{Total sum of squares} = \sum (Y_n - \bar{Y})^2 = \text{SST}$$

Sum of squares of predicted $\hat{Y}_n =$

"explained" sum of squares =

$$\sum (\hat{Y}_n - \bar{Y})^2 = \text{SSE}$$

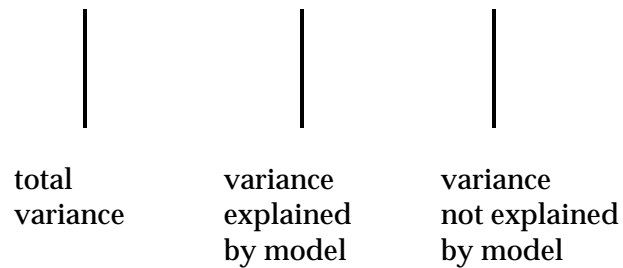
$$R^2 = \frac{\text{SSE}}{\text{SST}}$$

Example: $R^2 = .73$ means 73% of variance in y is explained by the model.

Another way to view R^2

$$Y_n = \hat{Y}_n + r_n$$

$$\text{Var}(Y_n) = \text{Var}(\hat{Y}_n) + \text{Var}(r_n)$$



$$\text{SST} = \text{SSE} + \text{SSR}$$

$$R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST} - \text{SSR}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$$

|
portion of variance
that is unexplained