

# Least-squares regression

## Cautions about correlation and regression

Outline:

- Least-squares regression.
  - Equations of regression line: slope, intercept
  - Residuals and residual plot
  - Outliers and influential observations
- Cautions about correlation and regression

## Least-Squares Regression

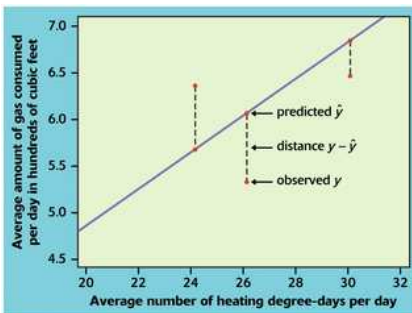
Regression describes the relationship between two variables in the situation where one variable can be used to explain or predict the other.

The regression line is a straight line that describes how a response variable  $y$  changes as an explanatory variable  $x$  changes.

1

### Fitting the Regression Line to Data

Since we intend to predict  $y$  from  $x$ , the errors of interest are mispredictions of  $y$  for a fixed  $x$ .



The **least-squares regression line** of  $y$  on  $x$  is the line that minimizes sum of squared errors.

This is the **least squares criterion**.

Given pairs of observations  $(x_1, y_1), \dots, (x_n, y_n)$ , the regression line is given by

$$\hat{y} = a + bx$$

where  $b = r \frac{s_y}{s_x}$  and  $a = \bar{y} - b\bar{x}$ .

3

2

### Interpreting the Regression Model

- The response in the model is denoted  $\hat{y}$  to indicate that these are predicted  $y$  values, not the true observed  $y$  values. The “hat” denotes prediction.
- The slope of the line indicates how much  $\hat{y}$  changes for a unit change in  $x$ .
- The intercept is the value of  $\hat{y}$  for  $x = 0$ . It may or not have a physical interpretation, depending on whether or not  $x$  can take values near 0.
- To make a prediction for an unobserved  $x$ , just plug it in and calculate  $\hat{y}$ .
- Note that the line need not pass through the observed data points. In fact, it often will not pass through any of them.

4

## Facts about Least Squares Regression

- The distinction between explanatory and response variables is essential. Looking at vertical deviations means that changing the axes would change the regression line.
- A change of 1 sd in  $x$  corresponds to a change of  $r$  sds in  $y$ .
- The least squares regression line always passes through the point  $(\bar{x}, \bar{y})$ .
- $r^2$  (the square of the correlation) is the fraction of the variation in the values of  $y$  that is explained by the least squares regression on  $x$ .

**When reporting the results of a linear regression, you should report  $r^2$ .**

These properties depend on the least-squares fitting criterion and are one reason why that criterion is used.

5

### Residual Plots

A **residual plot** is a scatterplot of the residuals against the explanatory variable. It can be used to assess the fit of the regression line.

Patterns to look for:

- *Curvature* indicates that the relationship is not linear.
- *Increasing or decreasing spread* indicates that the prediction will be less accurate in the range of explanatory variables where the spread is larger.
- *Points with large residuals* are outliers in the vertical direction.
- *Points that are extreme in the  $x$  direction* are potential high influence points.

**Influential observations** are individuals with extreme  $x$  values that exert a strong influence on the position of the regression line. Removing them would significantly change the regression line.

7

## Residuals

**Residuals** are the vertical distances between the data points and the corresponding predicted values.

$$\begin{aligned} r_i &= \text{observed } y - \text{predicted } y \\ &= y_i - \hat{y}_i \\ &= y_i - (a + bx_i) \end{aligned}$$

For a least squares regression, the residuals always have mean zero.

6

### A Regression Example

Consider the following data on unemployment rate and unemployment expenditure for several countries:

Country	Unemp. Rate	Unemp. Exp.
swz	0.5	0.16
lux	1.4	0.19
swd	1.6	0.72
jap	2.1	0.34
aut	3.3	0.88
fin	3.4	0.66
por	4.6	0.30
ger	4.7	1.17
nor	5.2	1.20
us	5.4	0.47
uk	6.8	0.88
gr	7.0	0.42
aus	7.0	1.01
bel	7.6	1.99
nl	7.8	2.25
nz	7.9	1.84
can	8.1	1.78
fr	8.9	1.34
den	9.7	3.22
it	10.3	0.40
ir	13.8	2.79
sp	15.9	2.43

### Summary Statistics

$$\bar{x} = 6.5$$

$$\bar{y} = 1.20$$

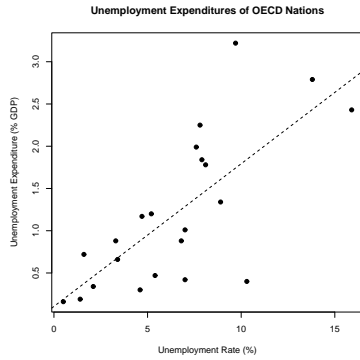
$$s_x = 3.87$$

$$s_y = 0.89$$

$$r = 0.73$$

8

## Regression Example (cont.)



**Regression Coefficients**

$$b = r \frac{s_y}{s_x}$$

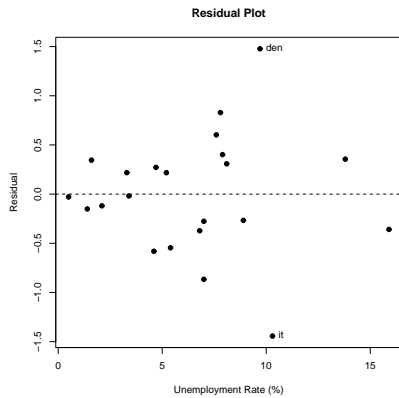
$$= 0.73 \frac{0.89}{3.87}$$

$$= 0.168$$

$$a = \bar{y} - b\bar{x}$$

$$= 1.20 - 0.168 \times 6.5$$

$$= 0.108$$



9

## Cautions about Correlation and Regression (cont.)

**Lurking variables** are variables not among the explanatory or response variables in a study that may influence the interpretation of relationships among the measured variables.

Lurking variables may falsely suggest a relationship when there is none, or may mask a real relationship.

**Association is not causation!** Two variables may be correlated because both are affected by some other (measured or unmeasured) variable.

### Example

Nations with more TV sets have higher life expectancies. Do TVs cause longer life? What's the real explanation?

Even if there is a causal relationship, it only makes sense in one direction. Sometimes the direction is obvious (e.g. if there is a time lag), but not always. For example, the high correlation between self esteem and success in school or work.

11

## Cautions about Correlation and Regression

- Correlation and regression describe only linear relationships.
- They are not resistant.

**Extrapolation** is the use of a regression line for prediction far outside the range of  $x$ -values used to obtain the line. Such predictions are not to be trusted.

**Averaging Data** smoothes out fine-scale variation, leading to higher correlation. This phenomenon is called **ecological correlation**. Results obtained on averages *should not* be applied to individuals.

### Example

In the 1988 CPS, the correlation between income and education for men age 25-64 was about 0.4.

Grouping data into nine census regions, averaging each variable within each region, and computing the correlation of the nine points yields  $r \approx 0.7$ .

10

## Establishing Causal Relationships

The best way to establish a causal relationship is to conduct an experiment where values of one (or several) variables are manipulated and the effect on some outcome is observed.

What if an experiment is not possible? There may be evidence for a causal relationship if:

- The association is strong
- The association is consistent across multiple studies.
- Higher doses are associated with stronger responses
- The alleged cause precedes the effect in time
- The alleged cause is plausible (perhaps because of “similar” studies, such as on animals)

12