Two Variables

We often wish to compare two different variables

Examples: different tests results, age and ability, education (in years) and income, speed and accuracy,...

Methods to compare two (or more) variables:

- correlation coefficient
- regression analysis

Notate bene!

- numeric variables

Background

Terminology: we speak of CASES, e.g., Joe, Sam, . . . and VARIABLES, e.g. height (h) and weight (w). Then each variable has a VALUE for each case, h_j is Joe's height, and w_s is Sam's weight.

We compare two variables by comparing their values for a set of cases,

- h_j vs. w_j
- h_s vs. w_s
- etc.

Tabular Presentation

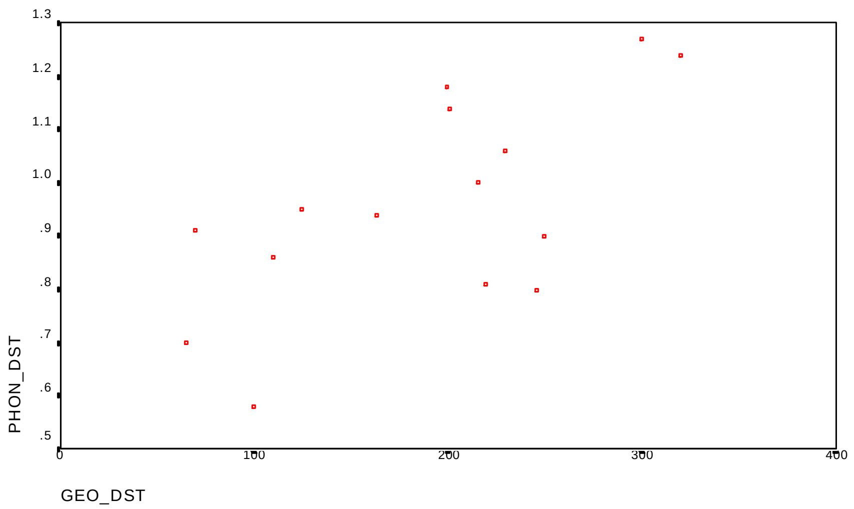
Example: Hoppenbrouwers measured pronunciation differences among pairs of dialects. We compare these to the geographic distance between places they're spoken.

Dialect Pair	Phon.Dist.	Geo.Dist.
Almelo/Haarlem	0.58	100
Almelo/Kerkrade	1.18	200
Almelo/Makkum	0.90	250
Almelo/Roodeschool	0.81	220
Almelo/Soest	0.91	70
Haarlem/Kerkrade	1.06	230
⋮	⋮	⋮
Kerkrade/Soest	1.14	201
Makkum/Roodeschool	0.95	125
Makkum/Soest	1.00	216
Roodeschool/Soest	0.94	163

Two variables—phonetic and geographic distance, and 15 cases (here, each pair is a separate CASE).

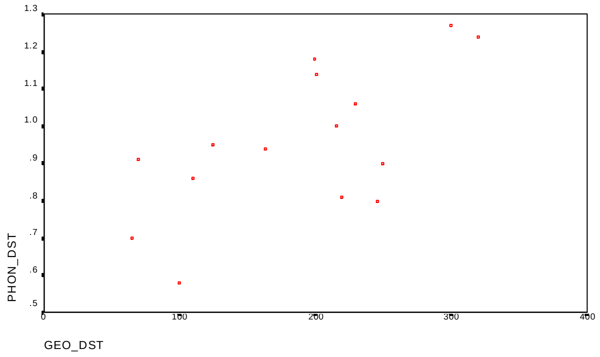
Scatterplots

One useful technique is to visualize the relation by graphing it.



Scatterplots

Each dot is a case, whose x-value is geo. distance, and y-value phon. distance.

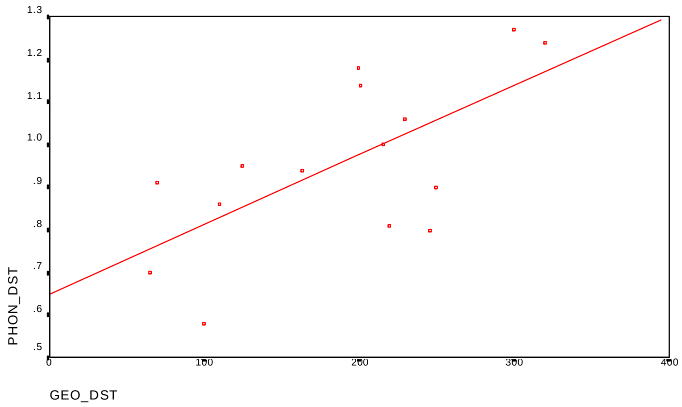


In general, we use x-axis for INDEPENDENT variables, and y for (potentially) DEPENDENT ones. We don't know whether phon. distance depends on geo. distance, but it might (while reverse is implausible).

Least Squares Regression

The simplest form of dependence is LINEAR—the independent variable determines a portion of the dependent value.

We can visualize this as fitting a straight line to the scatterplot.

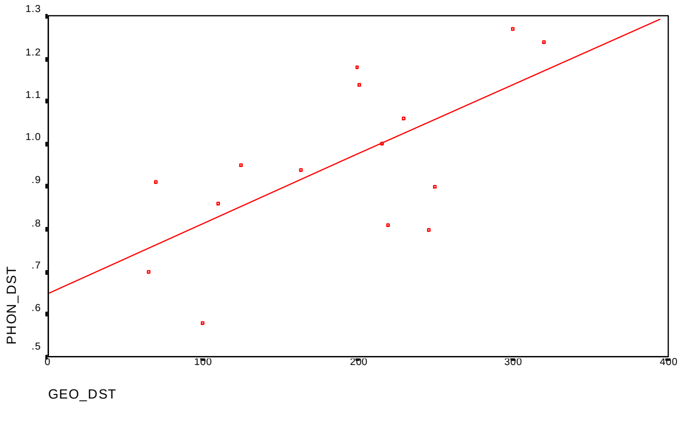


Least Squares Regression

If the scatterplot clearly suggests not a straight line, but rather a curve of another sort, you probably need to first TRANSFORM one of the data sets.

This is an advanced topic, but one which one need to keep in mind!

Least Squares Regression



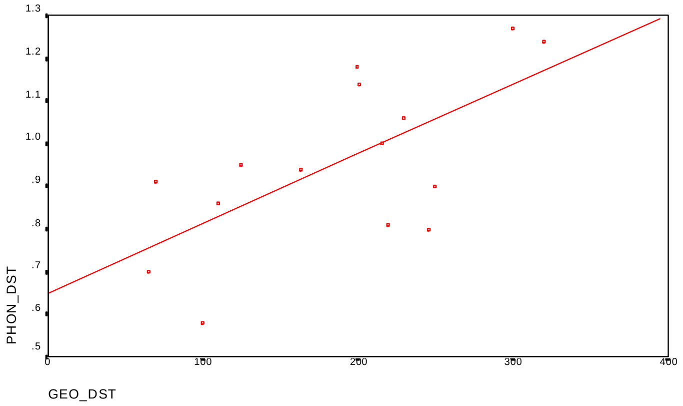
Like every straight line, this has an equation of the form: $y = a + bx$

a is the point where the line crosses the y-axis, the y-INTERCEPT, and b the SLOPE.

Predicted vs. Observed Values

The independent variable determines the dependent value (somewhat); this is the predicted value $\hat{y}$—the value on the line.

Note also the actual y—the data dot, not always the same.



Residuals

The difference between predicted and actual values $d_i = (\hat{y}_i - y_i)$ is the RESIDUAL—what the linear model does not predict. It is the vertical distance between the dot and the line.

LEAST-SQUARES REGRESSION finds the line which minimizes the squared residuals—for all the data.

$$\sum_i {d_i}^2 = \sum_i (\hat{y}_i - y_i)^2$$

SPSS Regression

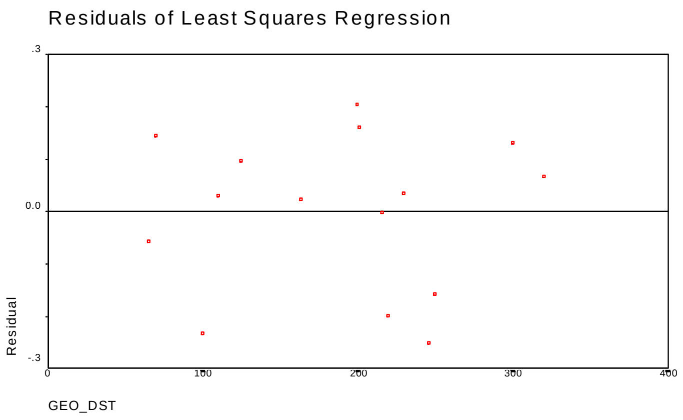
Least-sqares regression finds the best straight line which models the data (minimizes the squared error).

```
   * *   M U L T I P L E   R E G R E S S I O N   * *

Equation Number 1    Dependent Variable..   PHON_DST
Block Number  1.  Method:  Enter      GEO_DST

Analysis of Variance [ignore!]

----------- Variables in the Equation -------------
Variable               B         SE B

GEO_DST          .001631  5.1714E-04
(Constant)       .649778     .104898
```

$$y = 0.65 + 0.0016x$$

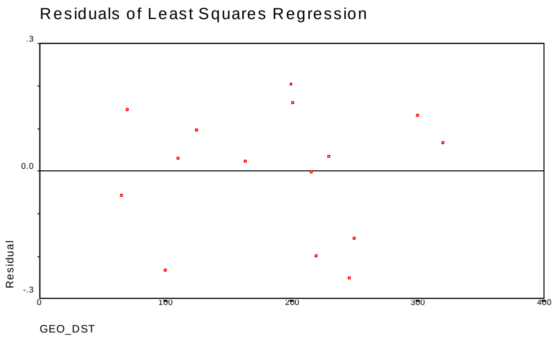
Residuals

Regression finds best line, but is sensitive to extreme values. Examine residuals.



Note requirement in regression model that residuals be normally distributed. Check via normal-quantile plot!

SPSS Plot of Residuals



Save residuals as new variable, then graph vs. original x value.

Watch out for extreme x values—influential, though residual may be small. See example 2.12 in Moore and McCabe.

Also examine OUTLIERS—large residuals.

Least Squares Regression*

(optional)

How does regression work?

We express the squared residuals as a function of the line. This is a function in two variables: a, the intercept, and b, the slope.

$$\begin{aligned}
f(a,b) &= \sum_i {d_i}^2 \\
&= \sum_i (\hat{y}_i - y_i)^2 \\
&= \sum_i ((a + bx_i) - y_i)^2 \\
&= \sum_i (a + bx_i - y_i)^2 \\
&= \sum_i a^2 + 2abx_i - 2ay_i + b^2{x_i}^2 - 2bx_iy_i + {y_i}^2
\end{aligned}$$

To minimize this function, find where its derivative $f' = 0$.

Least Squares Regression*

$$f(a,b) \quad = \quad \sum_i a^2 + 2abx_i - 2ay_i + b^2{x_i}^2 - 2bx_iy_i + {y_i}^2$$

To minimize a function in two variables, look at partial derivatives in ${f_a}'$, ${f_b}'$

$$\begin{array}{l} {f_a}'(a,b) = \sum_i 2a + 2bx_i - 2y_i \\ {f_b}'(a,b) = \sum_i 2ax_i + 2b{x_i}^2 - 2x_iy_i \end{array}$$

We then set each partial derivative to zero, and solve (the pair of linear equations).

Inf. Stats

Regression—Tiny Example*

Dialect Pair	Phon.Dist.	Geo.Dist.
Almelo/Haarlem	0.58	100
Almelo/Kerkrade	1.18	200
Kerkrade/Roodeschool	1.27	300

$$\begin{array}{rcl} {f_a}'(a,b) & = & \sum_i 2a + 2bx_i - 2y_i \\ & = & 2a + 2b(100) - 2 \times 0.58 + \\ & & 2a + 2b(200) - 2 \times 1.18 + \\ & & 2a + 2b(300) - 2 \times 1.27 \\ & = & 6a + 1200b - 6.06 \\ {f_b}'(a,b) & = & \sum_i 2ax_i + 2b{x_i}^2 - 2x_iy_i \\ & & 2a(100) + 2b(100)^2 - 2 \times 100 \times 0.58 \\ & & 2a(200) + 2b(200)^2 - 2 \times 200 \times 1.18 \\ & & 2a(300) + 2b(300)^2 - 2 \times 300 \times 1.27 \\ & = & 1200a + 280,000b - 1350 \end{array}$$

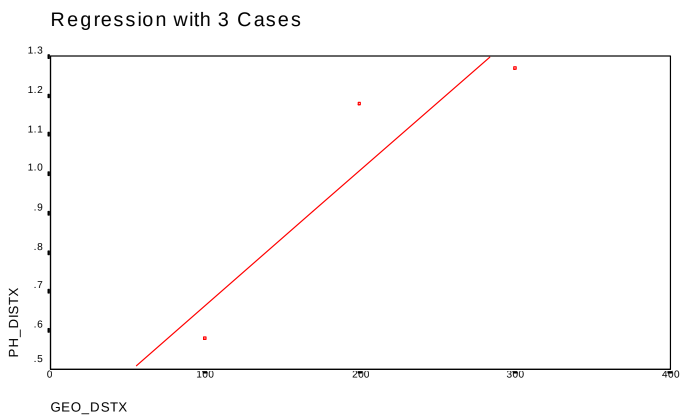
Regression—Tiny Example*

Now we solve these two linear equations (set to zero).

$$\begin{aligned}
0 &= 6a + 1200b - 6.06 \\
6a &= 6.06 - 1200b \\
a &= 1.01 - 200b \\
0 &= 1200(1.01 - 200b) + 280,000b - 1350 \\
&= 1212 - 240,000b + 280,000b - 1350 \\
40,000b &= 1350 - 1212 \\
b &= 138/40,000 = 0.00345 \\
a &= 1.01 - 1200(0.00345) = 0.32
\end{aligned}$$

Example, Cont.

$$y = 0.32 + 0.00345x$$



Example, Cont.


```
  * *   M U L T I P L E   R E G R E S S I O N   * *
Equation Number 1    Dependent Variable..  PH_DISTX
Variable(s) Entered on Step Number
   1..     GEO_DSTX

----------- Variables in the Equation ----------
Variable               B         SE B

GEO_DSTX        .003450     .001472
(Constant)      .320000     .318041
```


Linear Regression

- Asymmetric—appropriate when one variable might be "explained" by a second
  - Reaction time on basis of difficulty — negative!
  - Child's ability on basis of parents'
  - etc.
- No answer (yet) to how well does x explain y

CORRELATION analysis provides answers.

- Symmetric measure of extent to which variables predict each other
- Answer to how well does x explain y

Regression, correlation inappropriate when "best line" not straight (need transformations).

Correlation Coefficient

aka "Pearson's product-moment correlation"

$$r_{x,y} = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

- reflects strength of relation
  0 no correlation
  1 perfect positive correlation
  -1 perfect negative correlation
- no necessary dependence!
  shoe size, reading ability correlate—both dependent on age

Correlation Coefficient

Inf. Stats

```
    - -  Correlation Coefficients  - -

               GEO_DST     PHON_DST
GEO_DST        1.0000        .6584
              (    15)     (    15)
              P= .         P= .008

PHON_DST        .6584      1.0000
```

—geographic and phonetic distance correlate at 0.65

Correlation

$$r_{x,y} = \frac{1}{n-1} \sum \left(\frac{x - \overline{x}}{s_x} \right) \left(\frac{y - \overline{y}}{s_y} \right)$$

alternative:

$$r_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

- r “pure number” — no units
- insensitive to scale, percentages, ...
  corr. w. temperature can ignore scale
- symmetric $r_{x,y} = r_{y,x}$

Properties of Correlation

$$r_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}$$

- r measures “clustering” relative to σ_x, σ_y
  as $r \rightarrow 1$(or -1), dots cluster near regression line
- careful when “eyeing” data
  - change in σ affects apparent clustering
  - separate clusters lose in correlation
  - watch for nonlinear relations

Correlation/Regression

regression analysis —r^2: how much of y's variance may be atributed to x?

- regression requires that residuals be normally distributed
- nonsymmetric: y analyzed as dependent on x
- smoothed plot of y averages (for x groups)
- always flatter than SD line, the line with slope σ_y/σ_x which passes through (m_x, m_y)
- regression line (Gauss):

$$y = a + bx$$

then

$$b = r\frac{\sigma_y}{\sigma_x}$$

Interpretation of Correlation via Averages.

Example: height, weight have corr. coeff. $r = 0.5$

$\mu_h = 178$cm, $\mu_w = 72$kg, $\sigma_h = 6$cm, $\sigma_w = 6$kg

- for each σ_x, there are $r \cdot \sigma_y$'s
- what is ave. weight of those 184 cm tall?

$$\begin{aligned}
184\text{cm} &= 178 + 6\text{cm} \\
&= \mu_h + 1 \cdot \sigma_h \\
\delta_{\sigma_h} &= 1 \\
\overline{w}_{184}\text{cm} &= \mu_w + r_{w,h} \cdot \delta_{\sigma_h} \cdot \sigma_w \\
&= 72\text{kg} + 0.5 \cdot 1 \cdot 6\text{kg} \\
&= 75\text{kg}
\end{aligned}$$

Interpretation of Correlation via Averages.

- for each σ_x, there are $r \cdot \sigma_y$'s, $0 \leq r \leq 1$.
- $\Rightarrow \delta_w$ less (in z terms) than δ_h

since $r \leq 1$ averages of correlated variables must "regress" toward mean

Regression Fallacy

Height/weight example: In figuring $\overline{w}$ for restricted groups: δ_w less (in z terms) than δ_h

Since $r \leq 1$, averages of corr. var. **must** "regress"—but this is purely mathematical, no causal

"Regression fallacy": —seeing causation in regression

- height correlation between parents and children ($r = 0.4$)
  but very tall parents have less tall children (still taller than ave.)
- test-retest situations show extremes (high and low) closer to mean on second test
- "the course showed no general improvement, but the worst students improved"

Correlation

- measures strength of linear relation
- symmetric $r_{x,y} = r_{y,x}$
- related to slope of regression line

Caution needed:

- outliers — reduce r
- nonlinear association, e.g. intensity vs. loudness
- "ecological correlations" use averages, rates
  popular in politics, but overstate r (based on individuals)
- correlation $\not\to$ causation
  example: shoe size and reading ability

Regression Error

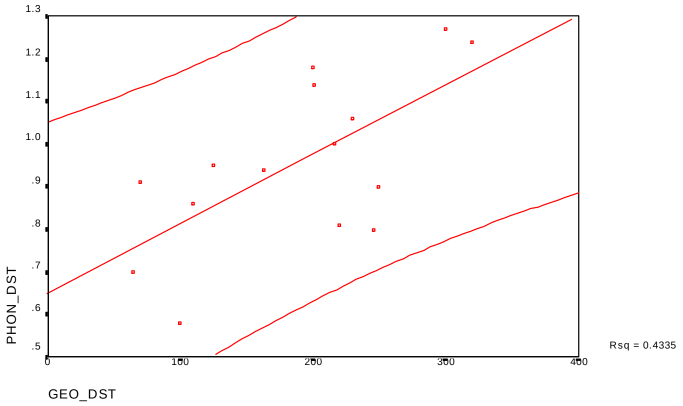
$\mu : \sigma$ :: regression line : regression error

regr. error measures dispersion around regr. **line**

regr. error can be calculate as standard deviation (from regression line, but also (shorter) $= \sqrt{1 - r^2} \times \sigma_y$

n.b. reg.error $\leq \sigma_y$

SPSS Plot of Regression Error



Shows ± 2 standard errors around regression line—where 95% of data must be found.

Next: Multiple Regression