Statistiek II

John Nerbonne

Information Science, Groningen
j.nerbonne@rug.nl
Slides improved a lot by Harmut Fitz, Groningen!

March 24, 2010

Correlation and regression

We often wish to compare two different variables

Examples: compare results on two distinct tests

- age and ability
- education (in years) and income
- speed and accuracy

Methods to compare two (or more) variables:

- Correlation coefficient
- Regression analysis

Notice:

- Correlation and regression only for numeric variables!

Background

Terminology: we speak of

- **cases**, e.g., Joe, Sam, etc. 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 versus w_j
- h_s versus w_s
- etc.

Tabular presentation

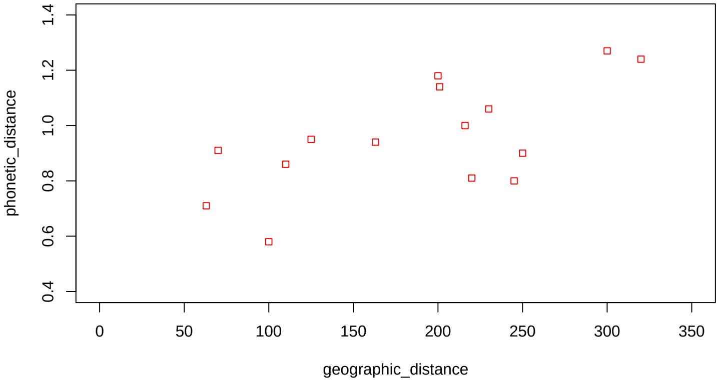
Example: Hoppenbrouwers measured pronunciation differences among pairs of dialects. We compare these to the geographic distance between places where they are spoken.

Dialect pair	Phon. dist.	Geogr. 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/Kerkrad	1.06	230
⋮	⋮	⋮
Kerkrade/Soest	1.14	201
Makkum/Rodeschool	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

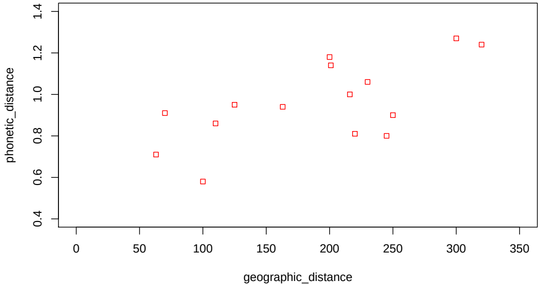
One useful technique is to visualize the relation by graphing it:



Scatterplot shows the relationship between two quantitative variables

Scatterplots

Each dot is a case, whose x-value is geographic distance, and y-value is phonetic distance.

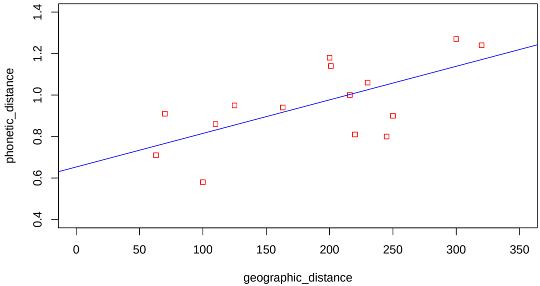


In general, we use x-axis for **independent** variables, and y-axis for **dependent** ones. We don't know whether phonetic distance depends on geographic distance, but it might (while reverse is implausible).

Least squares regression

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

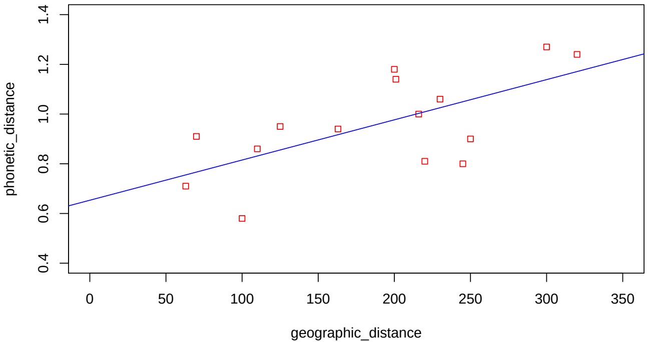
We can visualize this by fitting a straight line to the scatterplot:



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 something to keep in mind!

Least squares regression



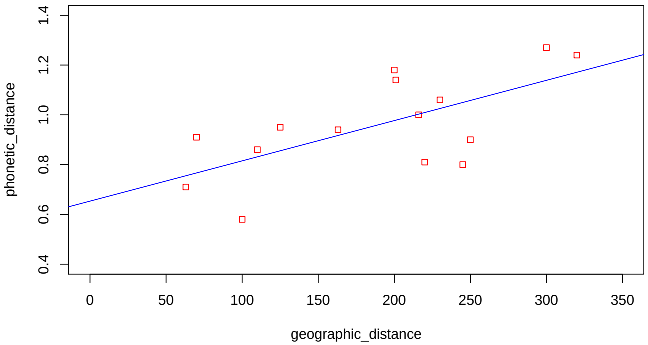
Like every straight line, this has an equation of the form:
$y = a + bx$

a is the point where the line crosses the y-axis, the **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 that the actual value y—the data dot—is not always the same as $\hat{y}$



Residuals

The difference between observed and predicted values

$$\epsilon_i := (y_i - \hat{y}_i)$$

is the **residual**—what the linear model does not predict. It is the vertical distance between the data point and the regression line.

Least-squares regression finds the line which minimizes the squared residuals—for all the data:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Regression with R

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

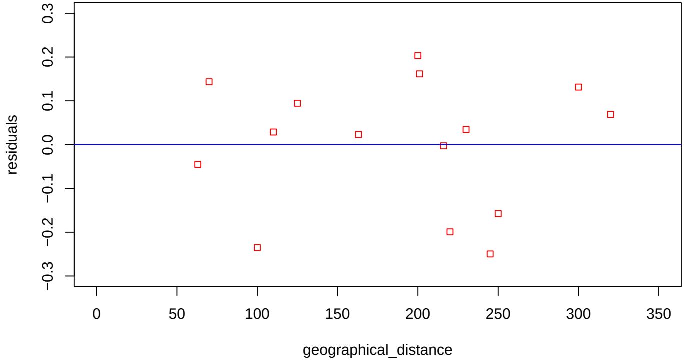
```
Call:
lm(formula = phonetic distance ~ geographic distance)

Coefficients:
            (Intercept)   geographic distance
            0.653292      0.001618
```

Regression line: $y = 0.65 + 0.0016x$

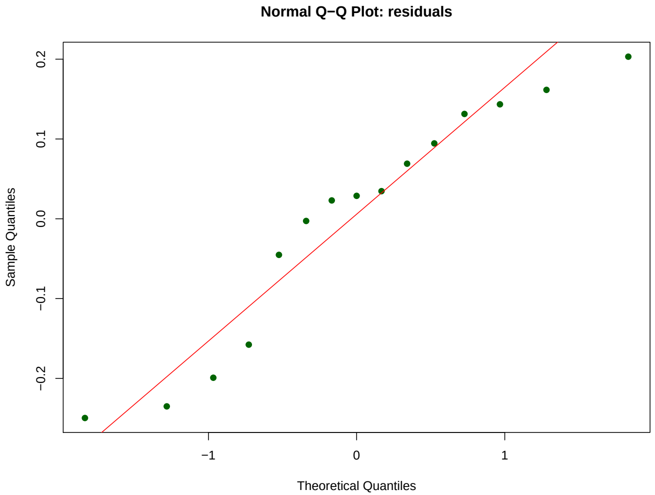
Residuals

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



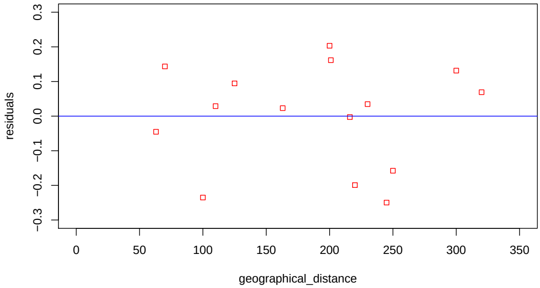
Note: requirement in regression model that residuals be normally distributed. Check with normal QQ-plot!

Check normality of residuals



Residuals look reasonably normal (Shapiro-Wilk test $p = 0.18$)

R plot of residuals



Save residuals as new variable, then graph against original x value

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

Also examine **outliers**—large residuals.

Least squares regression

How does regression work?

Suppose we have a sample $\mathcal{S} = (x_i, y_i)$ with $i = 1, \ldots, n$.

Let $x := (x_1, \ldots, x_n)$ and $y := (y_1, \ldots, y_n)$

We want to **estimate the regression line** $y = a + bx$ for this data.

This amounts to optimizing the intercept a and slope b with respect to the residuals:

Find a and b such that for a given sample $\mathcal{S}$ the sum of squared residuals is minimized.

Estimating the regression line

We express the sum of squared residuals as a function of the (unknown) regression line:

$$\begin{aligned}\sum_{i=1}^{n} \epsilon_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - (a + bx_i))^2 \\ &= \sum_{i=1}^{n} (y_i - a - bx_i)^2 \\ &= \sum_{i=1}^{n} (a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)\end{aligned}$$

Thus, $\sum_{i=1}^{n} \epsilon_i^2$ is function f in x, y with unknown parameters a, b.

Estimating the regression line

For a fixed sample $\mathcal{S} = (x, y)$, we want to minimize $f_{ab}(x, y)$ with

$$f_{ab}(x, y) = \sum_{i=1}^{n}(a^2 + 2abx_i - 2ay_i + b^2x_i^2 - 2bx_iy_i + y_i^2)$$

To minimize this function, find a and b such that $f'_{ab}(x, y) = 0$.

Treat a and b as variables and find partial derivatives $\frac{\partial}{\partial a}f$, $\frac{\partial}{\partial b}f$

$$\frac{\partial}{\partial a}f = f'_{xyb}(a) = \sum_{i=1}^{n}(2a + 2bx_i - 2y_i)$$

$$\frac{\partial}{\partial b}f = f'_{xya}(b) = \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_iy_i)$$

Regression—tiny example

Dialect pair	Phon. dist.	Geogr. dist.
Almelo/Haarlem	0.58	100
Almelo/Kerkrade	1.18	200
Kerkrade/Roodeschool	1.27	300

- plug these sample values into partial derivatives
- set them to zero
- solve pair of linear equations

$$
\begin{aligned}
f'_{xyb}(a) &= \sum_{i=1}^{n}(2a + 2bx_i - 2y_i) \\
&= 2a + 2b \cdot 100 - 2 \cdot 0.58 + \\
&\quad 2a + 2b \cdot 200 - 2 \cdot 1.18 + \\
&\quad 2a + 2b \cdot 300 - 2 \cdot 1.27 \\
&= 6a + 1200b - 6.06
\end{aligned}
$$

Regression—tiny example

$$
\begin{aligned}
f'_{xya}(b) &= \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_iy_i) \\
&= 2a \cdot 100 - 2b \cdot (100)^2 - 2 \cdot 100 \cdot 0.58 + \\
&\quad 2a \cdot 200 - 2b \cdot (200)^2 - 2 \cdot 200 \cdot 1.18 + \\
&\quad 2a \cdot 300 - 2b \cdot (300)^2 - 2 \cdot 300 \cdot 1.27 \\
&= 1200a + 280.000b - 1350
\end{aligned}
$$

Set to zero and solve:

$$
\begin{aligned}
& 0 = 6a + 1200b - 6.06 && \text{(I)} \\
\Leftrightarrow\ & 0 = a + 200b - 1.01 \\
\Leftrightarrow\ & a = 1.01 - 200b
\end{aligned}
$$

Regression—tiny example

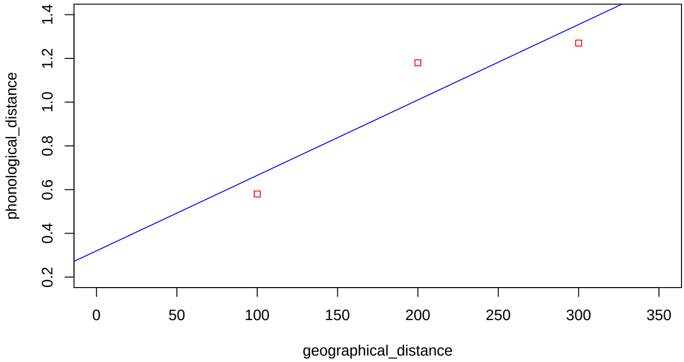
$$a = 1.01 - 200b \qquad \text{(I)}$$
$$0 = 1200a + 280.000b - 1350 \qquad \text{(II)}$$

Substitute *a* in (II) by (I):

$$\begin{aligned}
& 0 = 1200 \cdot (1.01 - 200b) + 280.000b - 1350 \\
\Leftrightarrow \quad & 0 = 1212 - 240.000b + 280.000b - 1350 \\
\Leftrightarrow \quad & 40.000b = 1350 - 1212 \\
\Leftrightarrow \quad & b = \frac{138}{40.000} = \underline{0.00345} \\
\Rightarrow \quad & a = 1.01 - 200 \cdot 0.00345 = \underline{0.32}
\end{aligned}$$

Hence, the regression line $y = 0.32 + 0.00345x$ minimizes the sum of squared residuals

Check calculations with R



Call:
lm(formula = phonetic distance ∼ geographic distance)

Coefficients:

(Intercept)	geographic distance
0.32000	0.00345

Linear regression

- Regression is asymmetric—appropriate when one variable might be ‘explained’ by a second
  - Reading times on the basis of difficulty—negative!
  - Child’s ability on the basis of parents’ ability
  - Final grade based on class attendance, etc.
- No answer (yet) to how well does x explain y
  Correlation analysis provides an answer
- Correlation symmetric measure of extent to which variables predict each other
- Answer to how well does x explain y

Regression and correlation inappropriate when ‘best fit’ is not straight line (need data transformations)

Correlation coefficient

How do you know if you are going to do well in a stats course?

Suppose you spend a lot of time on the material—more than your average class mate—then you'll have a high z-score in the distribution of study time.

You know that, generally, study time predicts grades.

So you know that you should have a high z-score in the distribution of grades.

If your final grade is not so good, I would expect you didn't spend much time studying. You would be below the mean in both distributions and have negative z-scores.

Correlation coefficient

If $x = (x_1, \ldots, x_n)$ is study time, and $y = (y_1, \ldots, y_n)$ are grades, we can measure correlation between the two variables as

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i}$$

- compute everyone's z-score (study time and grades)
- multiply both z-scores and sum for everyone in class
- divide by the degrees of freedom (# students -1)

Note: positive sum results from multiplying two positive or negative z-scores for x and y (positive correlation)

Negative sum (correlation) results from multiplying positive and negative z-scores (and vice versa)

No correlation results from mixed-sign z-scores with sum close to zero.

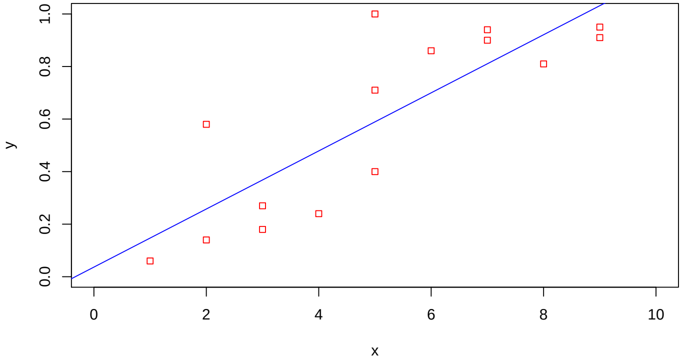
Correlation coefficient

Correlation coefficient aka "Pearson's product-moment coefficient"

$$r_{xy} = \tfrac{1}{n-1} \sum_{i=1}^{n} z_{x_i} \cdot z_{y_i} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma_x} \right) \left(\frac{y_i - \overline{y}}{\sigma_y} \right)$$

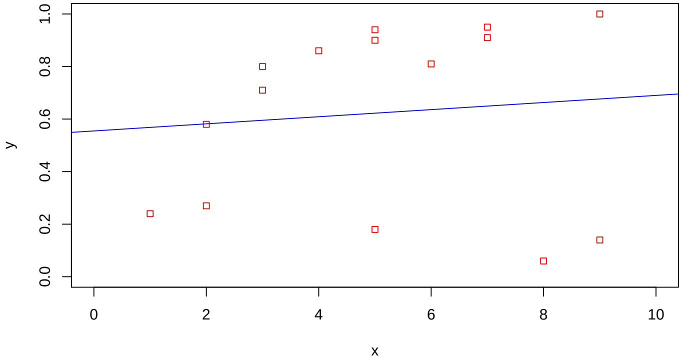
- r_{xy} reflects the strength of the relation between x and y
  - $r_{xy} = 0$ no correlation
  - $r_{xy} = 1$ perfect positive correlation (all data points on a straight line with positive slope)
  - $r_{xy} = -1$ perfect negative correlation
- no necessary dependence!
  - shoe size and reading ability correlate—both dependent on age

Visualizing correlation



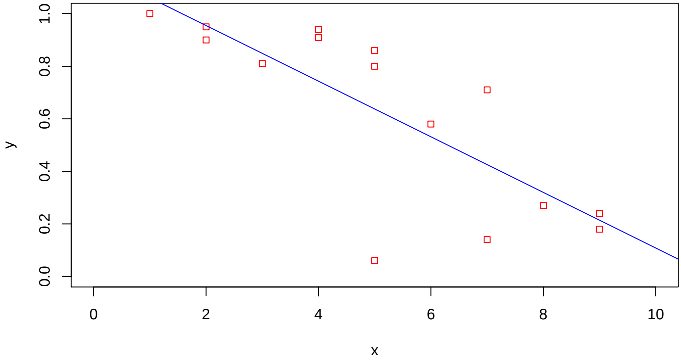
- data points lie close to the regression line
- correlation coefficient $r_{xy} = 0.83$
- strong positive correlation

Visualizing correlation



- data points scatter in a cloud around regression line
- correlation coefficient $r_{xy} = 0.1$
- no correlation (there might be correlation in both subsets)

Visualizing correlation



- data points close to regression line with negative slope
- correlation coefficient $r_{xy} = -0.77$
- correlation, but negative

Back to example: dialects

In our example: correlation coefficient for geographic and phonetic distance

In R simply call:

```
cor(phonetic-distance,geographic-distance)
[1] 0.6574452
```

Hence, phonetic and geographic distance correlate at $r = 0.66$

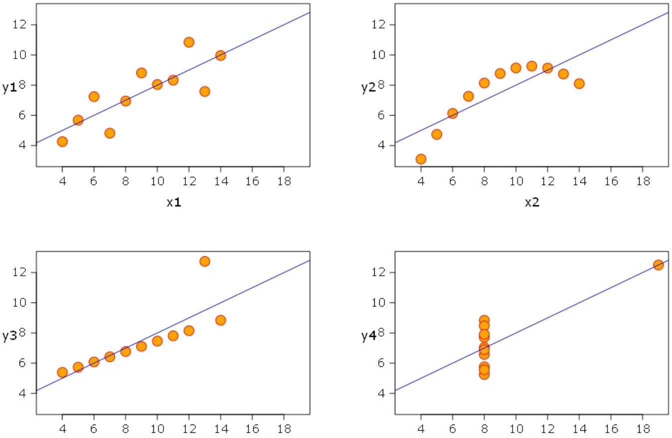
- r is a 'plain number'—no units
- insensitive to scale, percentages, etc.
  E.g., correlation with temperature can ignore scale (Celsius vs Fahrenheit)
- symmetric $r_{xy} = r_{yx}$

Properties of correlation

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma_x} \right) \left(\frac{y_i - \overline{y}}{\sigma_y} \right)$$

- correlation requires that both variables be quantitative (numerical)
- correlation coefficient always between 1 and -1
- as $r \rightarrow 1$ (or -1), dots cluster near regression line
- r measures 'clustering' relative to standard deviations σ_x, σ_y
- correlation can be misleading in the presence of outliers or nonlinear association
- therefore...

...always plot your data



Four variables *y* have same mean, standard deviation, correlation and regression line (examples from Anscombe)

Relationship between correlation and regression

Recall we obtained two partial derivatives (when minimizing sum of squared residuals):

$$f'_{xyb}(a) = \sum_{i=1}^{n}(2a + 2bx_i - 2y_i) \quad (1)$$

$$f'_{xya}(b) = \sum_{i=1}^{n}(2ax_i + 2bx_i^2 - 2x_iy_i) \quad (2)$$

Set (1) to zero:

$$\begin{aligned} & f'_{xyb}(a) = 0 \\ \Leftrightarrow \quad & n \cdot 2a + \sum_{i=1}^{n}(2bx_i - 2y_i) = 0 \\ \Leftrightarrow \quad & n \cdot 2a + 2b\sum_{i=1}^{n}x_i - 2\sum_{i=1}^{n}y_i = 0 \\ \Leftrightarrow \quad & n \cdot a = n \cdot \overline{y} - n \cdot b\overline{x} \\ \Leftrightarrow \quad & a = \overline{y} - b\overline{x} \end{aligned}$$

Relationship between correlation and regression

Plug $a = \overline{y} - b\overline{x}$ into (2) and set to zero:

$$
\begin{aligned}
& f'_{xya}(b) = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{n} (2(\overline{y} - b\overline{x})x_i + 2bx_i^2 - 2x_i y_i) = 0 \\
\Leftrightarrow \quad & (\overline{y} - b\overline{x})(n\overline{x}) + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \\
\Leftrightarrow \quad & n\overline{xy} - b\overline{x}^2 n + b\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0 \\
\Leftrightarrow \quad & b(\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n) = \sum_{i=1}^{n} x_i y_i - n\overline{xy} \\
\Leftrightarrow \quad & b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n}
\end{aligned}
$$

Relationship between correlation and regression

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - \overline{x}^2 n} \quad\Leftrightarrow\quad b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\Leftrightarrow\quad b = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\Leftrightarrow\quad b = \frac{1}{n-1} \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2\right)}$$

$$\Leftrightarrow\quad b = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{\sigma_x^2}$$

$$\Leftrightarrow\quad b = \left(\frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{\sigma_x}\right) \left(\frac{y_i - \overline{y}}{\sigma_y}\right) \right) \cdot \frac{\sigma_y}{\sigma_x}$$

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

Correlation and regression

Thus, the regression line $y = a + bx$ has

- slope $b = r\frac{\sigma_y}{\sigma_x}$ and
- intercept $a = \overline{y} - b\overline{x}$

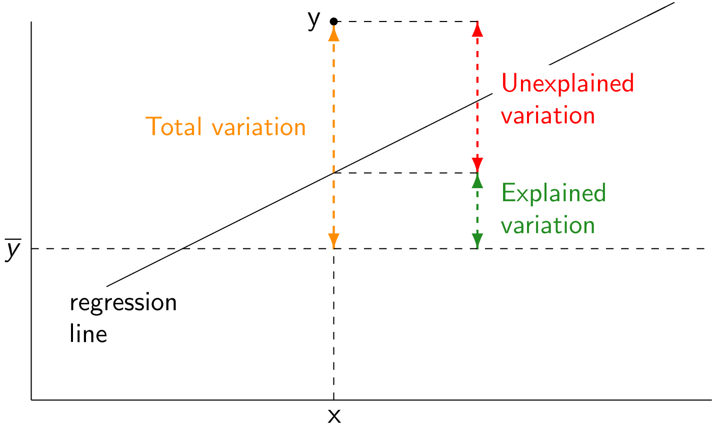
Consequently:

- correlation and regression are related via the coefficient r
- regression line always flatter than SD line, the line with slope $\frac{\sigma_y}{\sigma_x}$ which passes through $(\overline{x}, \overline{y})$

What's the point of regression analysis?

- analyze y as dependent on x (non-symmetric)
- determine how much of y's variance can be attributed to x

Correlation and regression



$$y - \overline{y} = (y - (a + bx)) + ((a + bx) - \overline{y})$$

Partitioning the variance

As in ANOVA, we can partition the variance in regression model:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \underbrace{(a + bx_i)}_{\text{regression line}})^2 + \sum_{i=1}^{n}(\underbrace{(a + bx_i)}_{\text{regression line}} - \overline{y})^2$$

Total variance = Unexplained variance + Explained variance

To what extent does explanatory variable x explain variation in response y? The quotient

$$\frac{\sum_{i=1}^{n}((a + bx_i) - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\text{explained variance}}{\text{total variance}}$$

measures this precisely.

Another relation between correlation and regression

$$\begin{aligned}
\frac{\text{explained variance}}{\text{total variance}} &= \frac{\sum_{i=1}^{n}((a+bx_i)-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \\
&= \frac{\sum_{i=1}^{n}((\overline{y}-b\overline{x}+bx_i)-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \\
&= \frac{\sum_{i=1}^{n} b^2(x_i-\overline{x})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2} \\
&= b^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2 \left(\frac{\sigma_y}{\sigma_x}\right)^2 \cdot \left(\frac{\sigma_x}{\sigma_y}\right)^2 \\
&= r^2
\end{aligned}$$

Coefficient of determination

$$\frac{\text{explained variance}}{\text{total variance}} = r^2 \qquad \text{("coefficient of determination")}$$

- r^2 indicates proportion of variability in data set that is accounted for by regression model
- provides a measure of how well future outcomes are likely to be predicted by the model
- in our example (phonetic distance of dialects):

$$r^2 = 0.66^2 = \underline{0.435}$$

Thus, 44% of the phonetic variation between dialects is accounted for by geographic distance

Interpretation of correlation via averages

Example: height, weight have correlation coefficient $r_{hw} = 0.5$

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

Slope of regression line: $b = r \cdot \frac{\sigma_w}{\sigma_h}$, i.e., for every σ_h, predicted weight changes by $r \cdot \sigma_w$

What is the average weight of those 184cm tall?

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

Regression toward the mean

In regression, for each σ_x, the predicted value of y changes by $r\sigma_y$

When there is less than perfect correlation, $0 \leq r < 1$

Hence, a predicted z_y for y will be closer to (the mean) 0 than z_x

In the previous example:

$$z_x = \frac{184\text{cm} - 178\text{cm}}{6\text{cm}} = 1, \; z_y = \frac{75\text{kg} - 72\text{kg}}{6\text{kg}} = 0.5$$

Since $r < 1$, averages of correlated variables **must** regress toward the mean ($z_y = r \cdot z_x$)

Regression toward the mean is a **mathematical inevitability**

Regression fallacy

Regression fallacy: seeing causation in regression

Examples:

(1) height correlation between parents and children ($r = 0.4$)

due to regression toward the mean, very tall parents tend to have less tall children (still taller than average)

Regression fallacy: tall father concludes his wife must have cheated

(2) motivation correlates with exam scores ($r = 0.5$)

test-retest situations show extremes (high and low scores) closer to mean on second test (regression toward mean)

Regression fallacy: bad students improved because I punished them

Correlation

Properties:

- only for numeric variables
- measures strength of a linear relation
- symmetric $r_{xy} = r_{yx}$
- related to the slope of the regression line

Caution needed:

- non-linear associations, i.e., curved patterns
- individual points with large residuals (outliers)
- influential observations (large deviation in x direction)
- "ecological correlations", i.e., correlations based on averages, popular in politics, overstate size of r
- correlation $\not\Rightarrow$ causation (e.g., shoe size and reading ability)

Inference for regression

Test whether regression yields significant association of variables:

Residual standard error: estimated standard error about the regression line

$$s = \sqrt{\frac{\sum_i^n e_i^2}{n-2}}$$

Standard error of the regression slope:

$$\mathrm{SE}_b = \frac{s}{\sqrt{\sum_i^n (x_i - \overline{x})^2}}$$

We test: $H_0 : b = 0, \quad H_a : b \neq 0$

Calculate t-statistic: $t = \frac{b}{\mathrm{SE}_b}$

Compare with critical t^* from $t(n-2)$

Inference for regression

In our example (phonetic variation in dialects):

$$s = \sqrt{\frac{0.3056}{13}} = 0.1533$$

$$\mathrm{SE}_b = \frac{0.1533}{\sqrt{298.2}} = 0.000514$$

$$t = \frac{0.001618}{0.000514} = 3.148$$

Critical value $t^* = 2.16$ (for t(df=13), $\alpha = 0.05$), hence reject H_0:

The data provides evidence in favor of a relationship between geographic and phonetic distance

Check with R

Call:
lm(formula = phonetic distance ~ geographic distance)

Residuals:

Min	1Q	Median	3Q	Max
-0.2496	-0.1015	0.0288	0.1129	0.2032

Coefficients:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.653292	0.104245	6.27	2.9e-05 ***
geographic distance	0.001618	0.000514	3.15	0.0077 **

—
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 .

Residual standard error: 0.153 on 13 degrees of freedom
Multiple R-Squared: 0.432, Adjusted R-squared: 0.389
F-statistic: 9.9 on 1 and 13 DF, p-value: 0.00773

Confidence intervals

What is the mean phonetic distance of dialects for $x^* = 150$km geographic distance?

$$\hat{y} = 0.65 + 0.0016 \cdot 150 = 0.89$$

Standard error for mean response $\hat{y}$ (for fixed x^*):

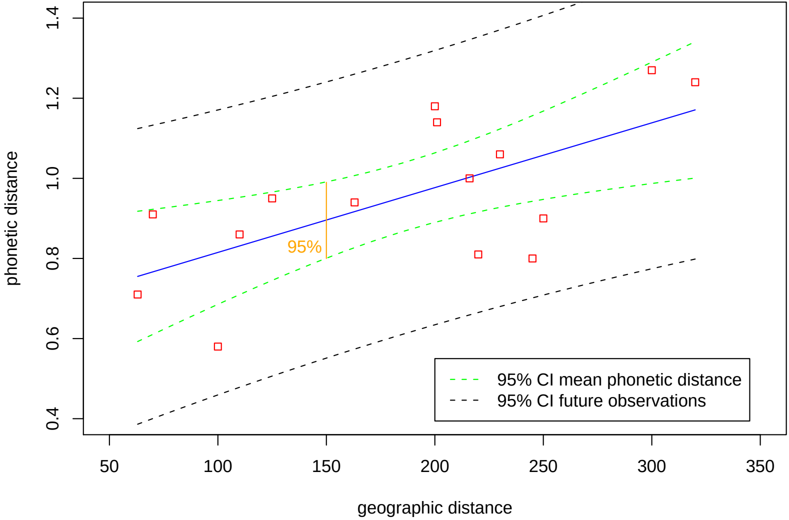
$$\text{SE}_{\hat{y}} = s \cdot \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{\sum_i^n (x_i - \overline{x})^2}}$$

Here: $\text{SE}_{\hat{y}} = 0.1533 \cdot \sqrt{\frac{1}{15} + \frac{(150-187.5)^2}{88914}} = 0.04403$

Confidence: $\hat{y} \pm t^* \text{SE}_{\hat{y}} = 0.89 \pm 2.16 * 0.04403 = 0.89 \pm 0.0951$

Hence, with 95% certainty, mean phonetic distance (for $x^* = 150$km) lies in the interval CI=(0.795,0.985)

Visualizing confidence intervals



Next week

Next week: multiple regression