Correlation and Regression

Idea: weight increases as height increase **or**
recognition time decreases as word frequency increases

- descriptive statistics
  - two or more numerical variables, e.g. height & weight, etc.
  - correlation: symmetric
  - regression: asymmetric
- may be interpreted inferentially
  - usually vs. H_0 “no relation”
    - correlation $H_0: \quad r = 0$
    - regresson $H_0: \quad m = 0$
      where m is slope of least-squares regression line

Correlation and Regression

- appropriate
  - two (or more) numerical measures on the **same** individuals
  - like paired t-test
    unlike χ^2, z-test, unpaired t-test
  - especially useful with two (or more) independent variables
  - example: incidence of heart attack (dependent)
    - amount of smoking (dependent $+$)
    - degree of overweight (dependent $+$)
    - frequency of physical exercise (dependent $-$)
    - ...

Reminder—Correlation

r – product of standardized values

$$r_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{s_x})(\frac{y_i-\bar{y}}{s_y}) \quad (1)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} z_{x_i} z_{y_i} \quad (2)$$

0	no correlation
1	perfect positive correlation
-1	perfect negative correlation

Reminder—Correlation

Inf. Stats

- r ranges: $-1 < r < 1$
- r “pure number” — no units
- insensitive to scale, percentages, ...
  correlation w. temperature can ignore scale
- symmetric $r_{x,y} = r_{y,x}$
  - no necessary **dependence**
  - shoe size and reading ability correlate in kids
    —both dependent on age
- r measures “clustering” relative to σ_y/σ_x
  as $r \rightarrow 1(\text{or } -1)$, dots cluster near line
- r sensitive to influential datapoints
  extreme x values

Example

A course in time management skills claims to completely change employees. You are sceptical, suspect that many skills are related to personality, experience, and custom. You obtain test scores given before and after the course to 25 employees. The test itself is regarded as reliable. Data:

5.8	6.0	5.9	6.1	5.7	6.0	5.9	6.1	5.7	5.9
5.2	5.1	5.6	5.9	5.9	5.9	5.8	6.0	6.1	6.3
5.9	6.3	6.9	7.0	6.3	6.4	5.9	6.0	5.1	6.2
5.7	6.1	6.0	6.2	5.1	5.0	6.1	5.9	6.4	6.0
6.2	6.1	5.8	6.0	5.6	6.4	6.8	7.1	6.3	6.9

$m_< = 5.9, \quad m_> = 6.1 \quad s_< = s_> = 0.45$

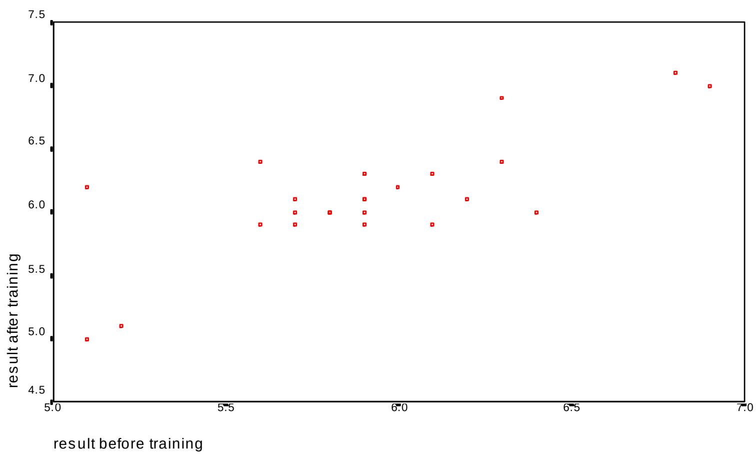
t-test possible, shows no sig. difference at $p = 0.01$

suspicion (H'): pre-course skills (most) important determinant of post-course skills

how to translate this into statistic?

Scores Correlate?

If pre-course results determine post-course results, they should **correlate**, i.e. $r \neq 0$



looks like positive tendency, calculate r

Scores Correlate?

H_0: $r = 0$ (no correlation in pre-, post-course skills)
H_a: $r \neq 0$ (correlation in pre-, post-course skills)

calculate r

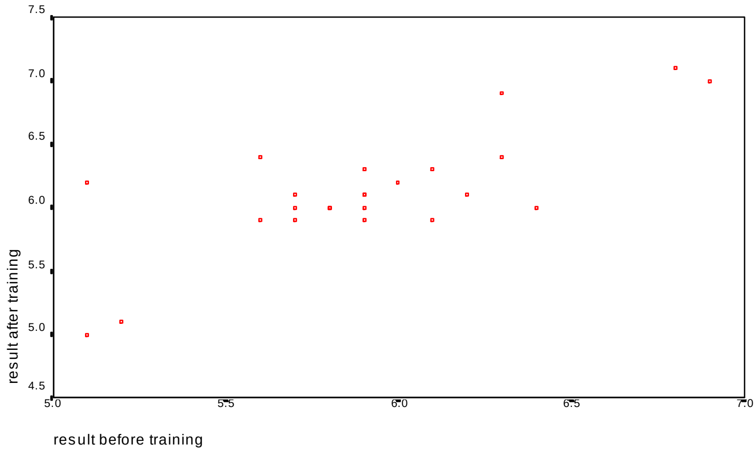
```
           - -  Correlation Coefficients  - -

                  AFTER        BEFORE

  AFTER           1.00         .77
                 (   25)      (   25)
                 P= .         P= .000

  BEFORE           .77        1.00
                 (   25)      (   25)
                 P= .000      P= .
```

result: certain correlation, reject H_0

no confidence interval calculated (complex)

Scores Correlate?

$r = 0.77, \quad p < 0.001$—but take care: correlation sensitive to influential data—look!



No apparent extremes, but recall connection correlation & regression.

Regression Analysis

We use regression to predict one numerical variable using another. Regression produces values b_0, b_1 in equation:

$$\hat{y} \;=\; b_0 + b_1 \times x$$

H_0: $b_1 = 0$ (no correlation in pre-, post-course skills)
H_a: $b_1 \neq 0$ (correlation is real)

Regression, too, tests hypothesis whether pre-course scores influence post-course scores. Can we predict post-course scores using pre-course scores?

Regression Analysis

Inf. Stats

Invoke linear regression to obtain

```
--- Variables in the Equation ---

Variable          B     ...

BEFORE          .81     ...
(Constant)     1.35     ...
```

$$\hat{y} \;=\; 1.35 + 0.81 \times x$$

But alone, this shows nothing about H_0!

Recall correlation-regression connection: $b_1 \;=\; r\frac{sy}{sx}$

Thus, tests closely related: $r = 0 \;\rightarrow\; r\frac{sy}{sx} = 0$

Regression Analysis

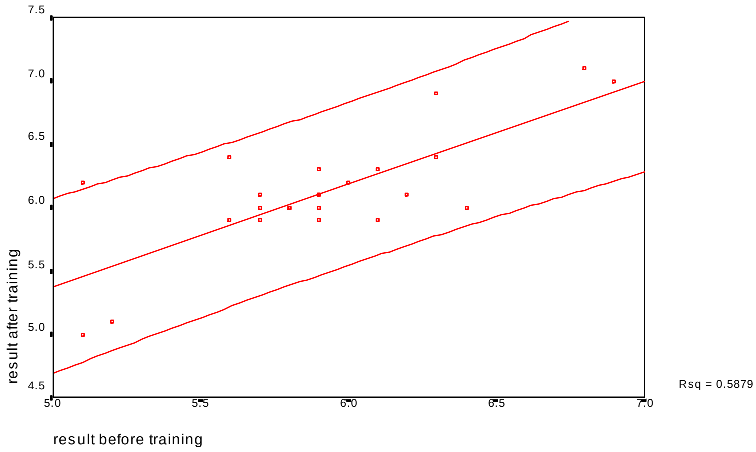
Two kinds of confidence intervals

- range of **individual** values (s_y)
- range of expected **means** (SE_y)

- Given x, where will 95% of corresponding y values be?
- Given x, where will 95% of corresponding samples means be?

Confidence Interval for Individuals

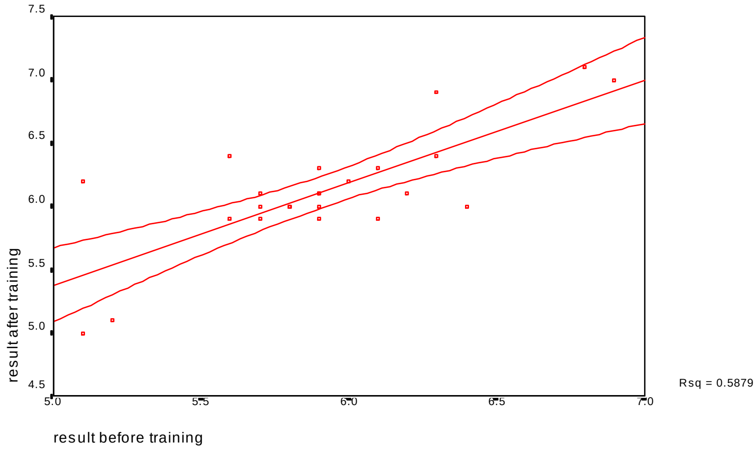
Invoke regression, confidence interval for **individuals**



Given x, where will 95% of corresponding y values be?

Confidence Interval for Means

invoke regression, confidence interval for **means**



Given x, where will 95% of corresponding y means be?

Confidence Interval for Means

—needed for hypothesis that $\beta_1 \neq 0$

two inferential steps possible

- derive confidence interval for β_1
- test $H_a: \quad \beta_1 \neq 0$

```
--- Variables in the Equation ---

Variable          B  ...   95% Confdnce Intrvl B

BEFORE          .81  ...      .52      1.1
(Constant)     1.35  ...     -.37      3.1
```

Given this sample, there is less than 2.5% chance that $\beta_1 < 0.52$. Pre-course scores are significant!

Alternative Test whether $H_a: \quad \beta_1 \neq 0$


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

Eq. Nr. 1    Dependent Variable..   AFTER   after training
Block Number  1.  Method:  Enter      BEFORE

Variable(s) Entered on Step Number
   1..     BEFORE    result before training
   .
   .
--- Variables in the Equation ---

Variable      B      SE B   95% Confdnce Intrvl  ...

BEFORE       .81     .14      .52     1.1
(Constant)  1.35     ...      ...     ...

----------- in ------------
Variable          T  Sig T

BEFORE        5.728  .0000
(Constant)    1.620  .1188
```

RuG

Methode behind Alternative Test

$$t \ = \frac{b_1}{s_{b_1}} = \frac{0.81}{0.14} = 5.7$$

Roughly, how many sd's is b_1 from 0?

$24 (= n - 1)$ degrees of freedom

$P(t(24) > 3.75) = 0.0005$ from table (hidden in SPSS)

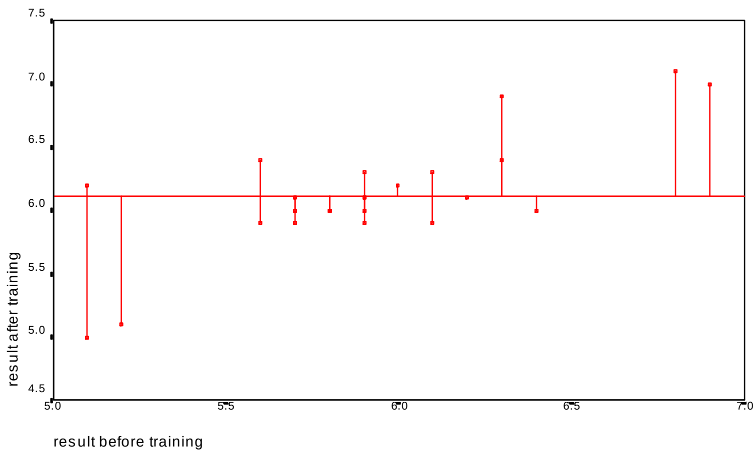
RuG

Examine Residuals!

Recall: regression sensitive to extreme x values, expects roughly normal distribution(s)

Check via residuals, invoke via SPSS regression analysis

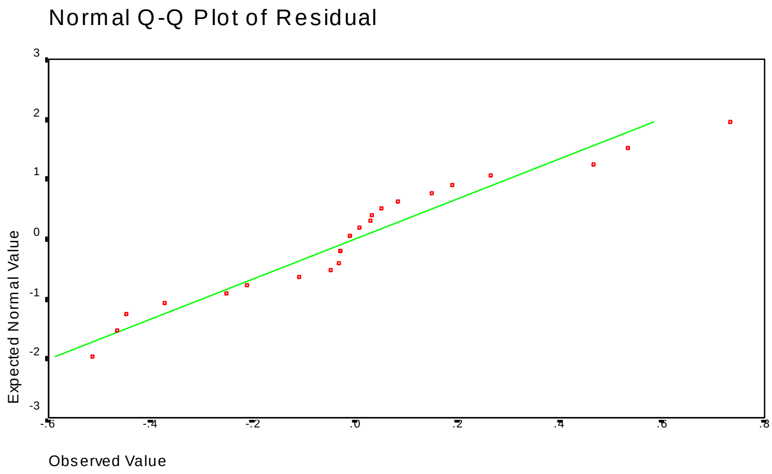
Residuals!



Still looks OK

Alternative view: normal quantile plots

Alternative View of Residuals: Normal Quantile Plots

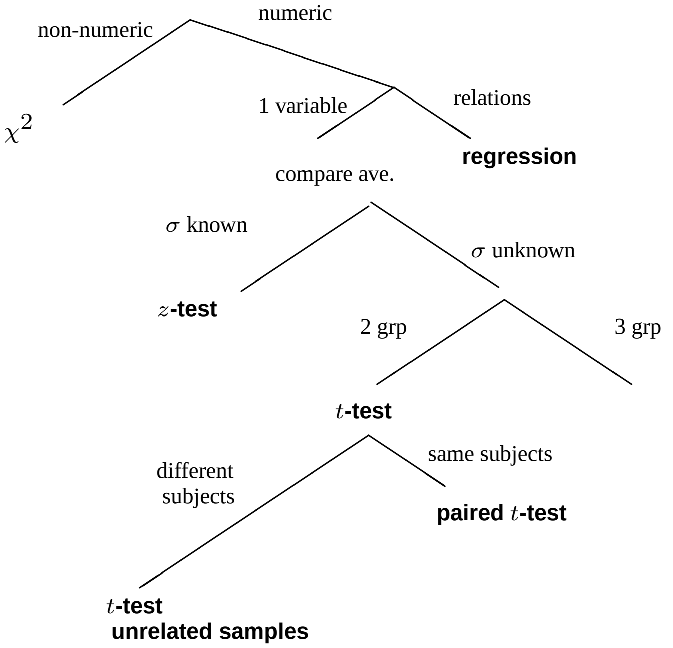


Normal distribution shows up clustered around straight line—this is fine.

Inf. Stats

Correlation/Regression

- Test for relation between two numeric variables
  - correlation: symmetric
  - regression: asymmetric x *influences* y
- Test contrasts vs. H_0 “no relation”
  - correlation $H_0: \quad r = 0$
  - regresson $H_0: \quad \beta_1 = 0$
    where β_1 is slope of least-squares regression line
- Assume normally distributed variables
- Extensions to multiple regression possible, very powerful


non-numeric
numeric
χ^2
1 variable
relations
regression
compare ave.
σ known
σ unknown
z-test
2 grp
3 grp
t-test
different subjects
same subjects
paired t-test
t-test unrelated samples

Residuals