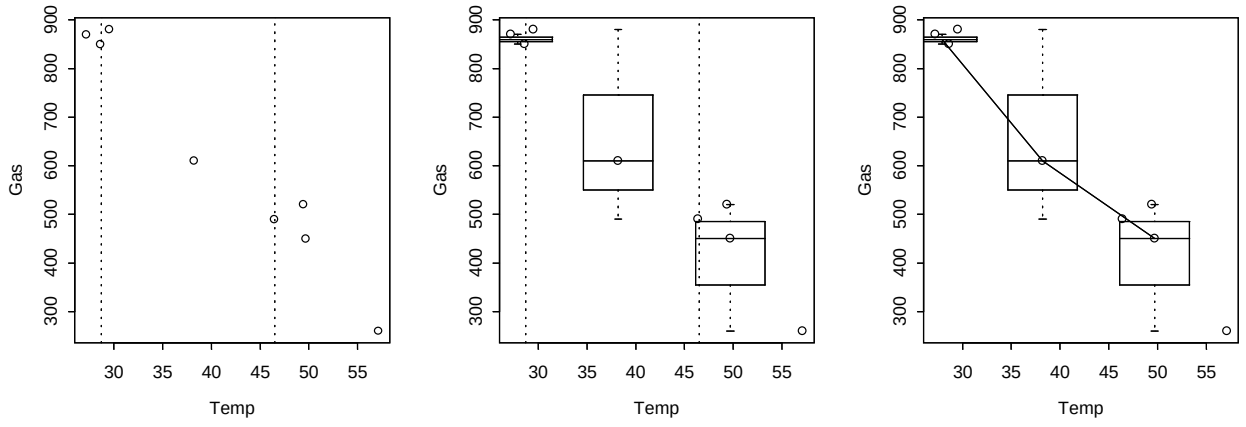
Quantifying the Relationship in a Scatter Plot

Illustrating a Trend: The Median Trace

- Divide the X-axis up into at least 3 equal-sized groups (be sensible!)
- Compute boxplots for the Y's in each group
- Connect the medians of the boxplots

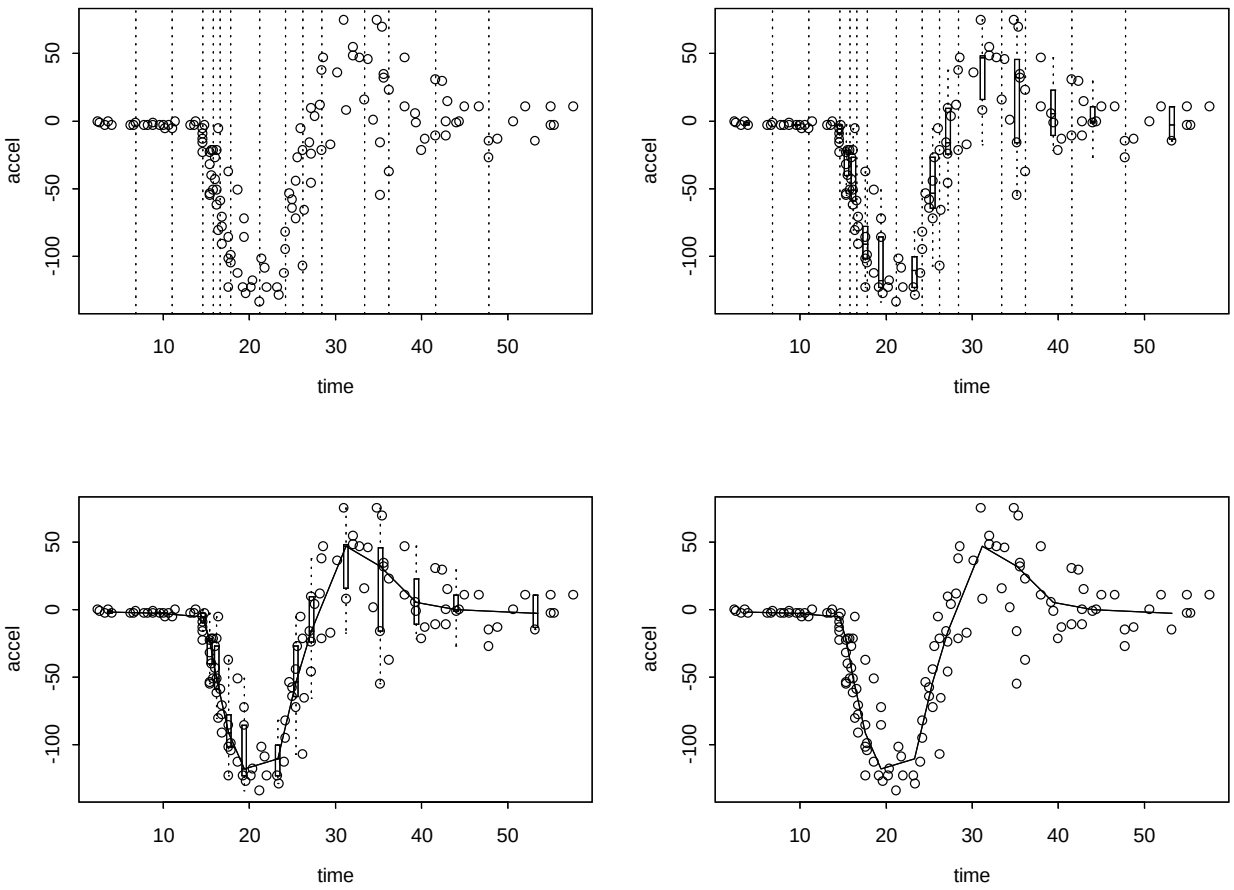
Example: Gas Consumption

- 8 observations
- 3 groups
- 2 or 3 observations per group



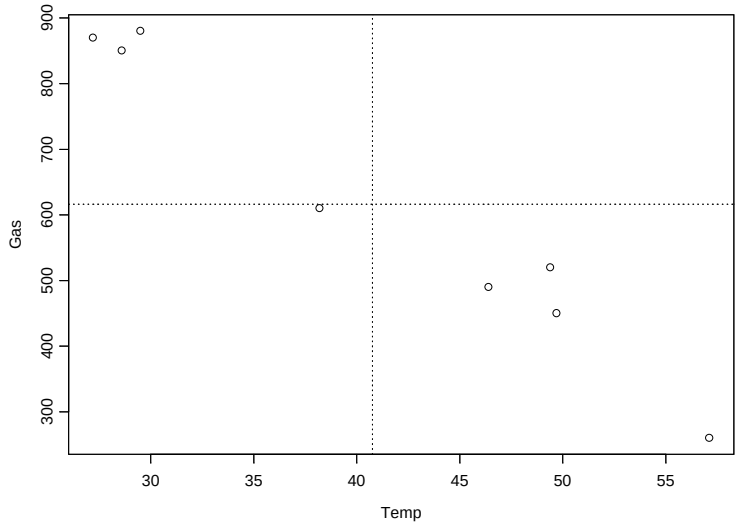
Example: Motorcycle Crashes

- 133 observations
- 15 groups
- 8 or 9 observations per group



Correlation: How much LINEAR relationship is there?

Example: Gas Consumption



	Mean	SD
Temp	40.76	11.45
Gas	616.25	229.41

Temp	Gas	Z for Temp	Z for Gas	Product
49.4	520	0.7540623	−0.41956313	−0.316376755
38.2	610	−0.2237088	−0.02724436	0.006094803
27.2	870	−1.1840197	1.10612098	−1.309669071
28.6	850	−1.0617983	1.01893903	−1.081907776
29.5	880	−0.9832275	1.14971195	−1.130428353
46.4	490	0.4921594	−0.55033605	−0.270853038
49.7	450	0.7802526	−0.72469995	−0.565449049
57.1	260	1.4262800	−1.55292846	−2.214910811

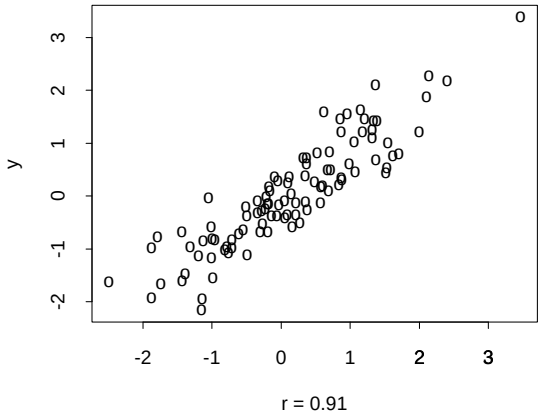
Correlation r = (Sum of products)$/(n-1) = -0.9833571.$

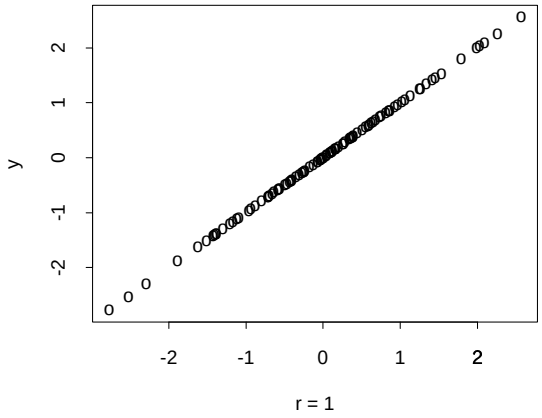
Correlation Rules of Thumb

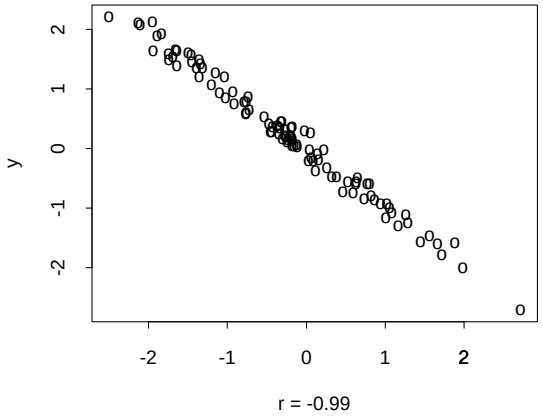
Correlation r must be between -1 and $+1$.

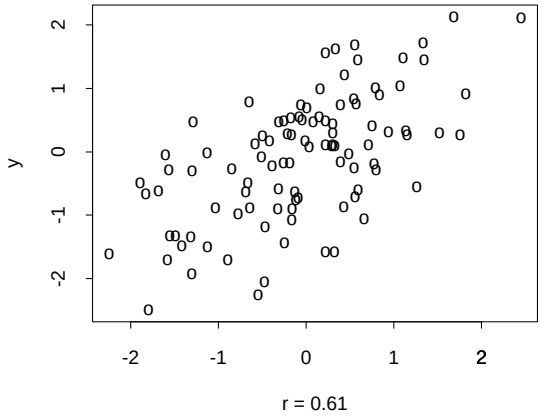
- Correlation r near $+1$: strong *increasing* linear trend
- Correlation r near -1: strong *decreasing* linear trend
- Correlation r near 0: no *linear* relationship

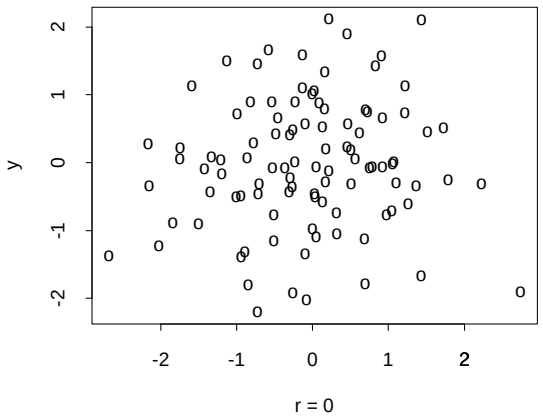
. . . but there might be a strong nonlinear relationship. . .

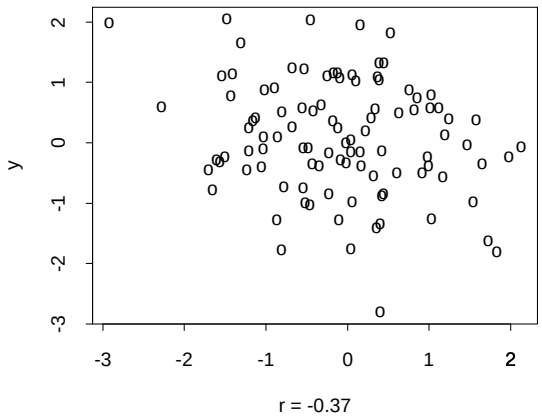

y
r = 0.91


y
r = 1

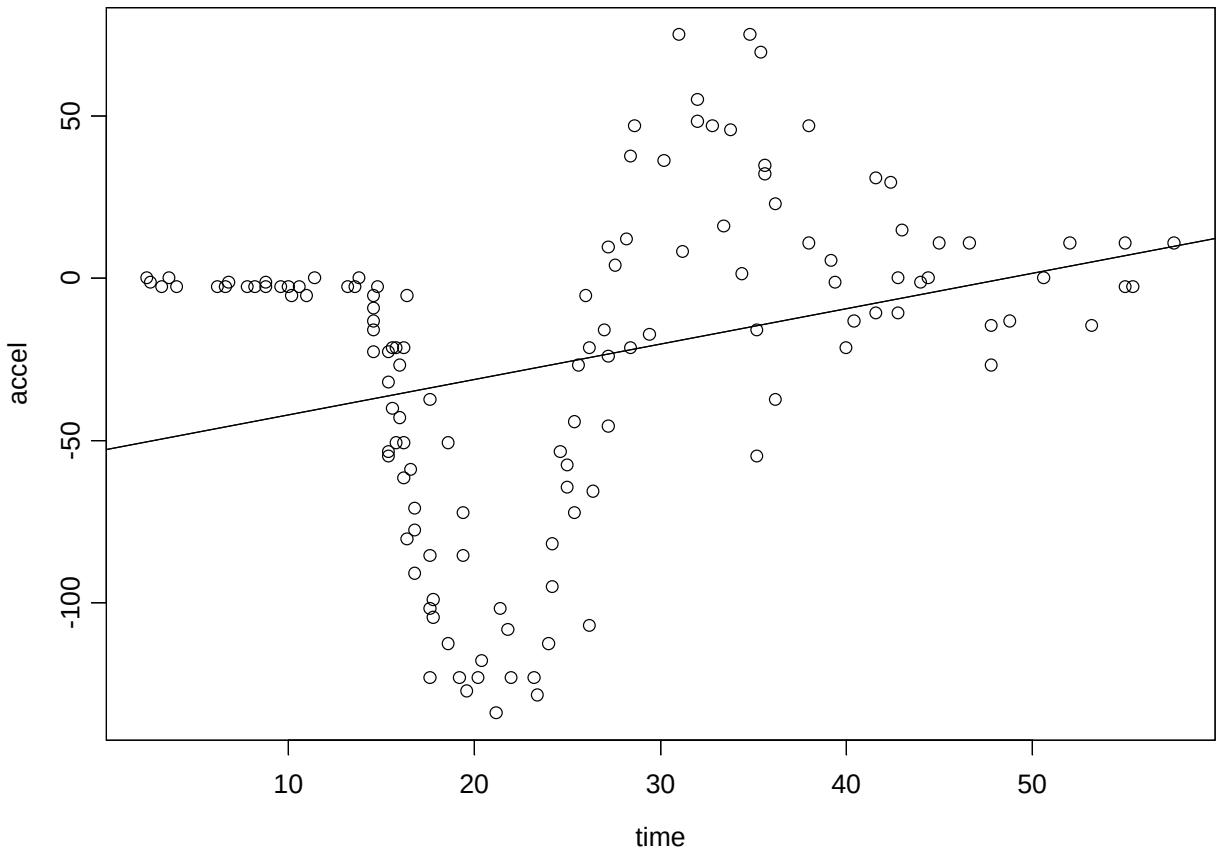

y
r = -0.99


y
r = 0.61


y
r = 0

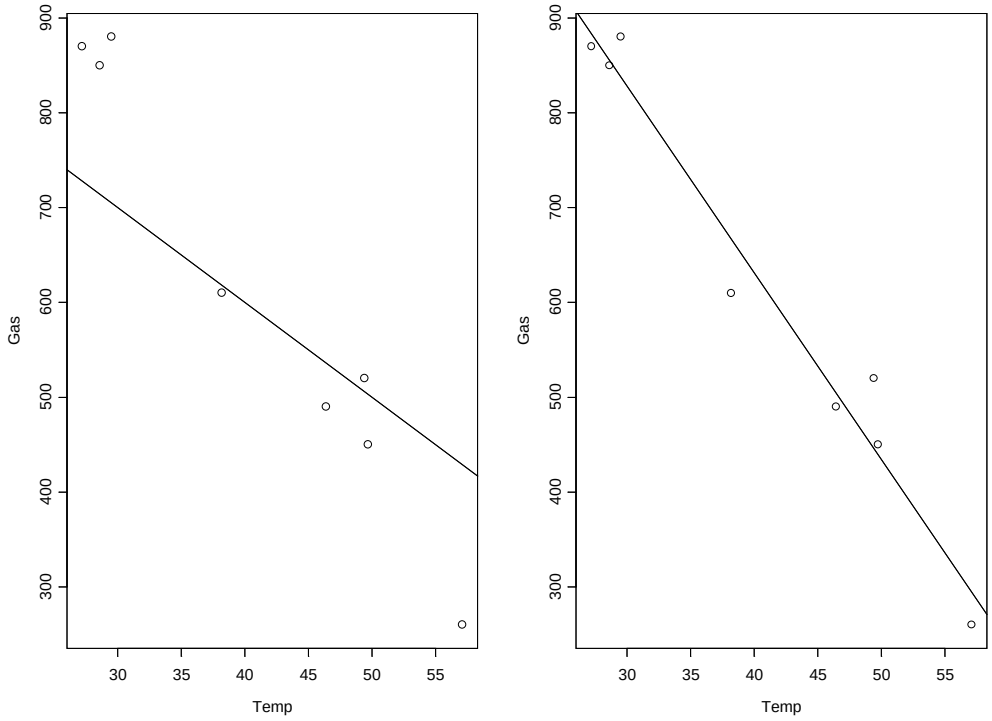

y
r = -0.37

Example: Motorcycle Crashes



- $r = 0.2964$
- Mild increasing relationship
- But correlation and *linear* relationships really don't describe the pattern of the data!

Equation of the “Best line”



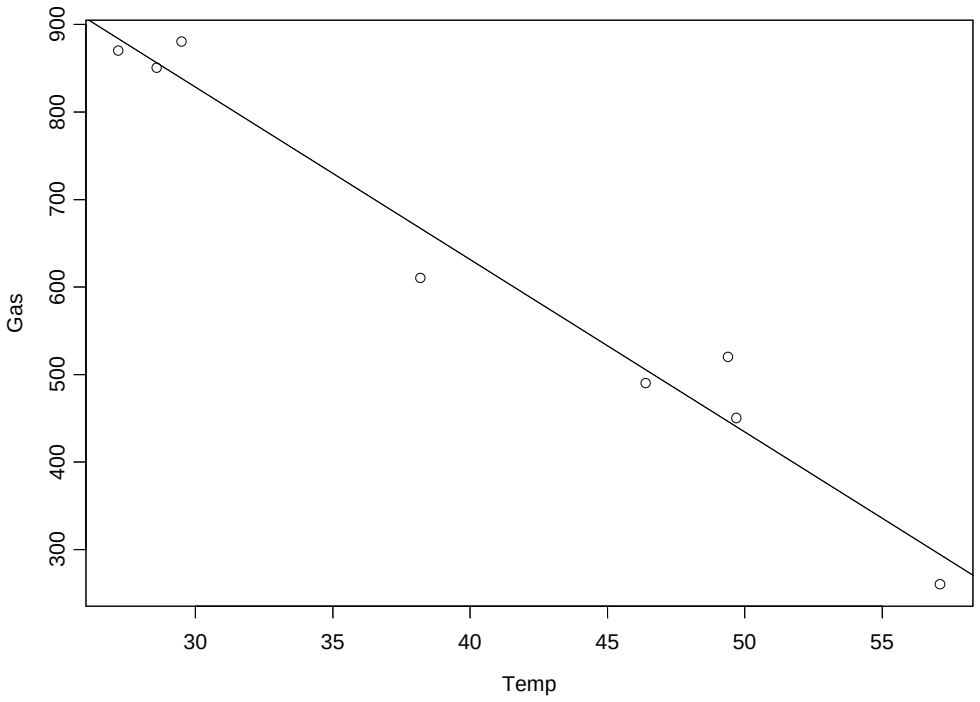
- Residuals: vertical distances from points to line
- The *least squares line* is the line that minimizes the sum of the squared residuals
- Equation of the least-squares line:

$$\begin{aligned} y &= a + b \cdot x, \text{ where} \\ b &= r \cdot \frac{SD_y}{SD_x} \\ a &= \text{mean}_y - b \cdot \text{mean}_x \end{aligned}$$

Put it all together:

$$y = \text{mean}_y + r \cdot \frac{SD_y}{SD_x}(x - \text{mean}_x)$$

Example: Gas consumption



		Mean	SD	r
X	Temp	40.76	11.45	–0.98
Y	Gas	616.25	229.41	

$$b = r \cdot SD_y/SD_x = (-0.98) \cdot 229.4111.45 = -19.66$$

$$a = \text{mean}_y - (-19.66) \cdot \text{mean}_x = 1417.59$$

$$y = 1417.59 - 19.66 \cdot x$$

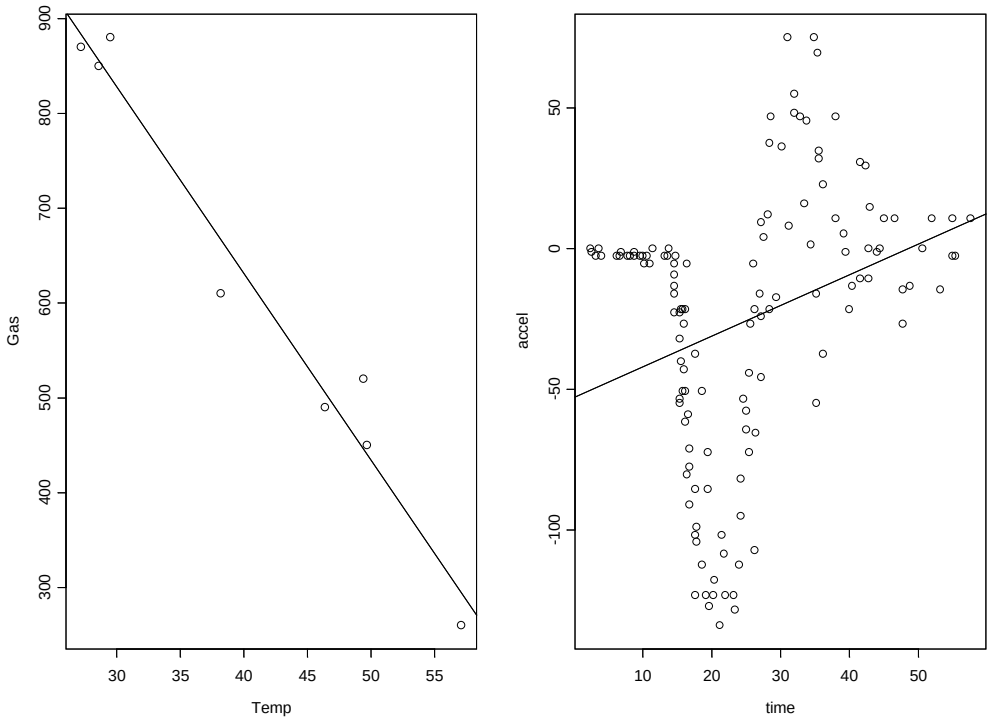
- If outside temp is $x = 40^o$F, then expect to use $y = 1417.59 - 19.66 \cdot 40 = 631.18$ Cubic Feet of gas.
- If outside temp is $x = 60^o$F, then expect to use $y = 1417.59 - 19.66 \cdot 60 = 237.98$ Cubic Feet of gas.

How good is a straight-line regression?

Quantitative answer:

> *r^2 is the fraction (or percent) of the variation in y that is explained as a linear function of x.*

- In the *Gas Consumption* data, $r = -0.98$, so $r^2 = 0.96$, or 96% of the variation in $y =$ Gas is explained as a linear function of $x =$ Temp.
- In the *Motorcycle Crash* data, $r = 0.30$, so $r^2 = 0.09$ or only about 9% of the variation in $y =$ accel is explained as a linear function of $x =$ time.



How strong in the relationship?

It depends on the field of study!

- In the physical sciences, $r \approx 0.9$ or better is *expected*; i.e. at least $r^2 = 80\%$ of the variability should be explained.

- In psychology and social sciences $r \approx 0.7$ is wonderful ($r^2 = 50\%$ of the variability explained) and $r^2 \approx 0.3$ is often acceptable (only $r^2 = 10\%$ of the variability explained!).