Topic 7: Linear Regression

Practice Test Questions: Correlation

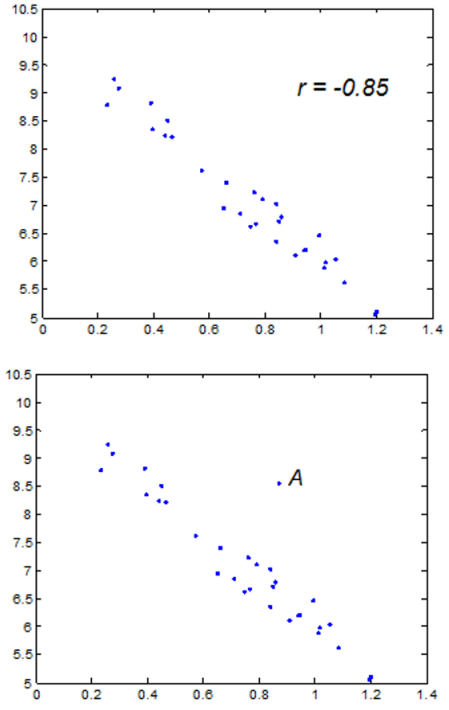
i	x_i	y_i	$(x_i - \bar{x})/s_x$	$(y_i - \bar{y})/s_y$	product
1	0	5	–1.1	+1.17	–1.28
2	2	2	***T***	–0.83	+0.25
3	3	2	+0.1	–0.83	–0.08
4	6	4	+1.3	+0.65	+0.65
	$\bar{x} = 2.75$	$\bar{y} = 3.25$			Sum = –0.46
	$s_x = 2.5$	$s_y = 1.5$			***r*** =

1. What is ***T*** ?
   A +0.3
   B –0.3
   C –0.75
   D 0.75
   E None of the above.

2. What is ***r*** ?
   A –0.46
   B –0.115
   C –0.153
   D 0
   E None of the above.

The two scatterplots show the same data, except that the point at *A* has been added to the dataset at the bottom. What can you say about the correlation coefficient of the dataset with *A* added?

A It is greater than –0.85.
B It is less than –0.85.
C It equals –0.85.
D It is impossible to tell without computing the correlation.



Least Squares Regression

A **regression line** describes ... ... how a response variable Y changes with respect to an explanatory variable X

We can use the relationship (according to this line) to make **predictions** about the behavior of Y

Regression line prediction: a hypothetical value for Y given a specific value for X

Unlike with correlation, it matters

which variable is Y (the response) and which variable is X (explanatory)

WARNING: a regression line should ONLY be used to predict response Y from explanatory X, not the other way around!

General equation of a line:

$$y = b_0 + b_1 x$$

y-intercept

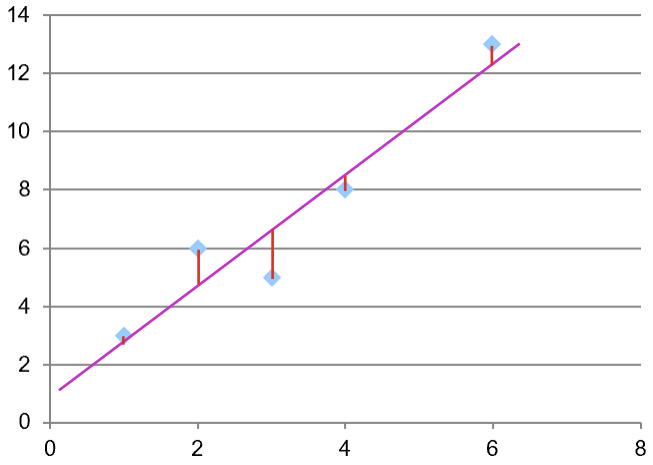
slope

Intercept interpretation: when x = 0, the value of y is $y = b_0$

Slope interpretation: when x **increases** by 1, the value of y **changes** by b_1

Small example:

X	Y
1	3
2	6
3	5
4	8
6	13



Is this line good? Best? How can we tell?

How to construct a line to fit these data?

General consensus:

The "line of best fit" is the line that MINIMIZES the SUM of the SQUARED vertical distances from each point to the line

In figure: red lines = vertical distances from points to line

Notation: Data is n points, each point has an x-value and a y-value $(x_1, y_1)\ (x_2, y_2) \ldots (x_n, y_n)$

Generic point: use subscript "i" such as (x_i, y_i)

$\hat{y}$: the value predicted for y by a regression line using some given x-value

Definition: A **residual**

Residual: the (vertical) difference between the OBSERVED value y_i for some datapoint and the PREDICTED value of y for the datapoint's x-value x_i

Prediction: $\hat{y}_i = b_0 + b_1 x_i$

Residual $= y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$

Mathematically, a least-squares regression line minimizes the sum

Least-squares regression line
= "line of best fit"
= line that minimizes SUM of SQUARED RESIDUALS

$$\sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_i)\right)^2$$

What values of b_0 and b_1 minimize this quantity?

$$b_1 = r \cdot \frac{s_Y}{s_X} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

r = correlation coefficient

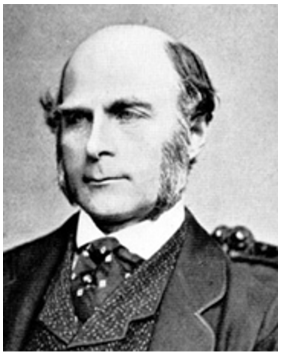
Sample means of y, x

s_y, s_x = sample standard deviations of y, x

Why the term "regression"?

Because 1 guy used the term in 1 specific study, and it stuck.

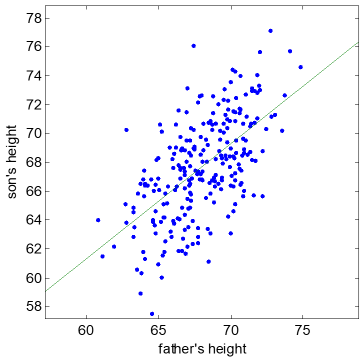
The term was coined by Francis Galton (1822 - 1911) in a study on the heights of fathers and sons. He found that fathers taller than average would more likely have shorter sons, and fathers shorter than average would more likely have taller sons. Galton called this "regression towards the mean height." (His 1886 paper was titled "Regression towards mediocrity in hereditary stature.")



Galton's data looked something like this:

For fathers taller than average, the least-squares line predicts a shorter height for their sons, and vice-versa. That is, the slope b_1 of the line is less than 1.

The name "regression line" stuck, even though it's only a "regression" when the units are the same on both axes and slope is less than 1.



Practice: Observations for explanatory variable X and response variable Y have

$$\bar{x} = 7, \quad \bar{y} = 5, \quad s_x = 4, \quad s_y = 2, \quad r = 0.85$$

Find the equation of a least-squares regression line that we could use to predict an observation of Y given an observation of X.

$b_1 = r\frac{s_y}{s_x}$
$= 0.85\left(\frac{2}{4}\right)$
$= 0.425$

$b_0 = \bar{y} - b_1\bar{x}$
$= 5 - (0.425)7$
$= 2.025$

Line:
$\hat{y} = 2.025 + 0.425x$

What if we knew the point (6,6) was in this dataset?

Prediction:
$\hat{y} = 2.025 + 0.425(6)$
$= 4.575$

Residual:
$y - \hat{y} = 6 - 4.575$
$= 1.425$

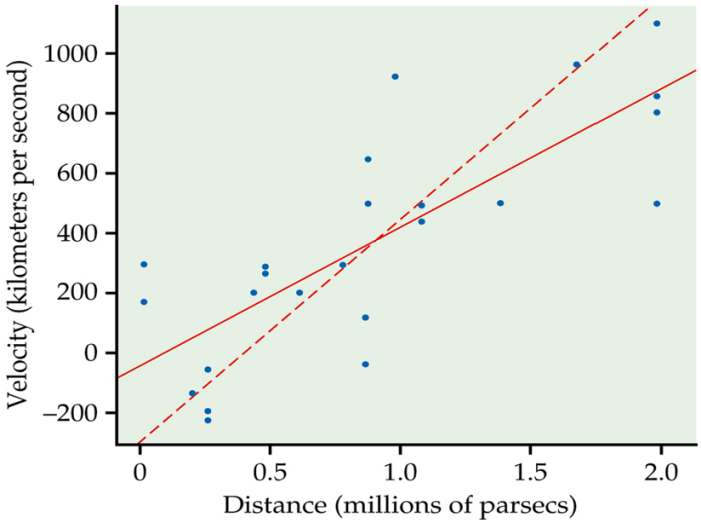
Interpretation:
Regression line UNDERESTIMATES (i.e. is SMALLER THAN) y by 1.425

Question: Will switching which variable is explanatory and which is the response change the least-squares line?

YES! In vast majority of cases, both slope and intercept will change when you switch explanatory and response.

Example: Data on galaxies

X: galaxy's distance from Earth
Y: velocity at which galaxy is moving away from Earth



Solid line: prediction of velocity for a given distance.

Dashed line: prediction of distance for a given velocity.

Connection between Correlation and Regression

The value r^2 measures ...

r^2 measures the percentage (i.e. proportion) of VARIATION observed in the response variable (Y) that is "explained" by the line of best fit

r^2 = square of correlation coefficient r

(and by the observed variation in X)

Rule of thumb:

Use r^2, not r, to describe STRENGTH of LINEAR relationship between X and Y

Still use r for DIRECTION (i.e. positive vs negative)

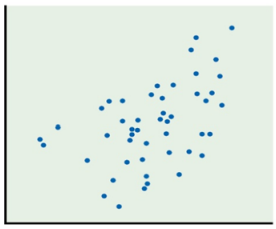
$r^2 = 0$

Correlation $r = 0$

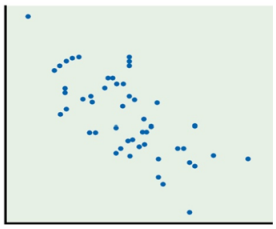
Correlation $r = -0.3$

$r^2 = 0.09$

$r^2 = 0.25$



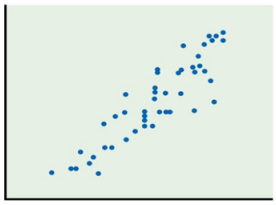
Correlation $r = 0.5$



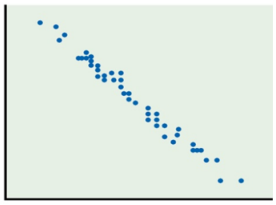
Correlation $r = -0.7$

$r^2 = 0.49$

$r^2 = 0.81$



Correlation $r = 0.9$



Correlation $r = -0.99$

$r^2 = 0.9801$

$r = 0.5 \implies r^2 = 0.25$

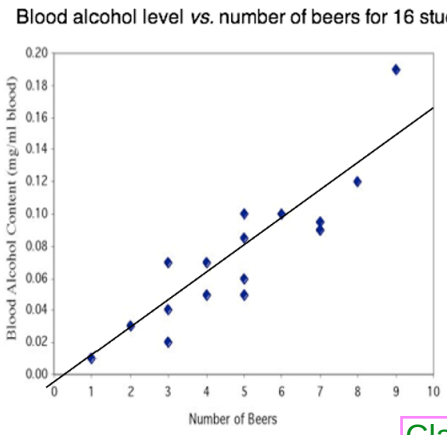
Interpretation: fairly weak (positive) linear relationship

$r = -0.7 \implies r^2 = 0.49$

Interpretation: moderate (negative) linear relationship

Limitations of using regression to make predictions

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05



Classic case of **extrapolation** error!

Regression equation:

$\hat{y} = -0.0127 + 0.018x$

Evaluate at x = 0: predicts y = -0.0127

BUT... y represents BAC, which can't be negative!

Extrapolation vs. interpolation:

Extrapolation: using a least-squares regression line to predict y-values based on x-values OUTSIDE the range of x-values in the data

Predictions based on **extrapolation** are considered BAD and UNTRUSTWORTHY!

Interpolation: using a least-squares regression line to predict y-values based on x-values INSIDE the range of x-values in the data

We always want our predictions to be INTERPOLATIONS

Practice test question: Researchers measured the two variables below for a large number of subjects.

X = sodium intake in mg per day, and
Y = systolic blood pressure mmHg

The least-squares regression line for predicting Y given X is

$$\hat{y} = -15.4 + 2.3x$$

Interpreting intercept...
Intercept = -15.4: predicted blood pressure for someone with 0mg/day sodium intake is -15.4mmHg
(this result is likely extrapolation error)

What is the predicted systolic blood pressure for someone whose sodium intake is 60 mg per day?

$\hat{y} = -15.4 + 2.3(60)$
$= 122.6$

Interpreting slope...
Slope = 2.3: increasing sodium intake by 1mg/day leads to 2.3mmHg increase in blood pressure

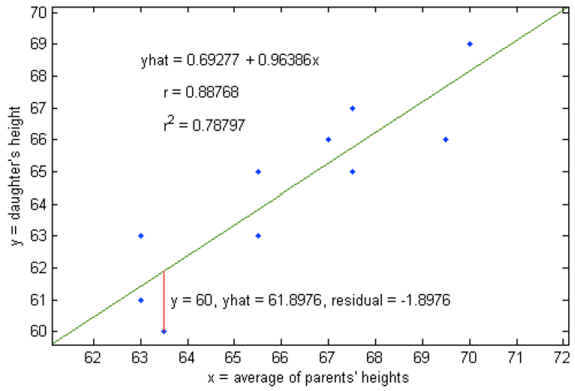
Residual Plots

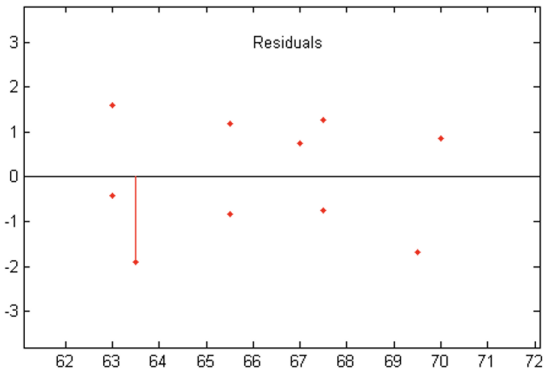
Start with a scatterplot of data + a least-squares regression line

Make a scatterplot where the y-values are the RESIDUALS of each datapoint!

Points: (x_i, y_i)

Line: $\hat{y} = b_0 + b_1 x$





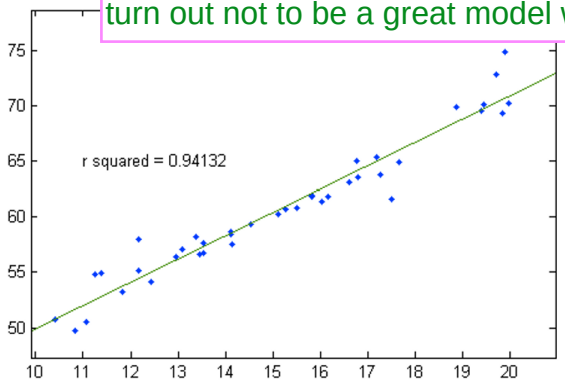
Points: $(x_i, y_i - \hat{y}_i)$

Line: $y_i - \hat{y}_i = 0$

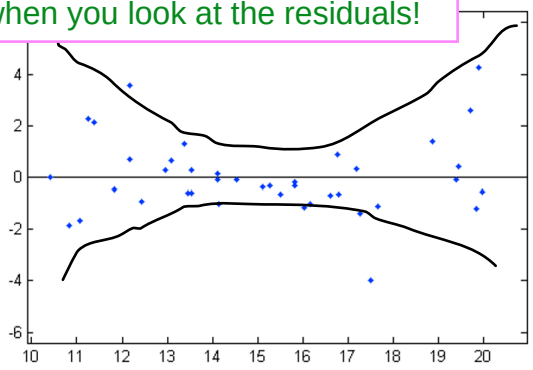
x-axis matches scatterplot x-axis

Why would we want to plot residuals by themselves?

A regression line that looks good in the regular scatterplot might turn out not to be a great model when you look at the residuals!



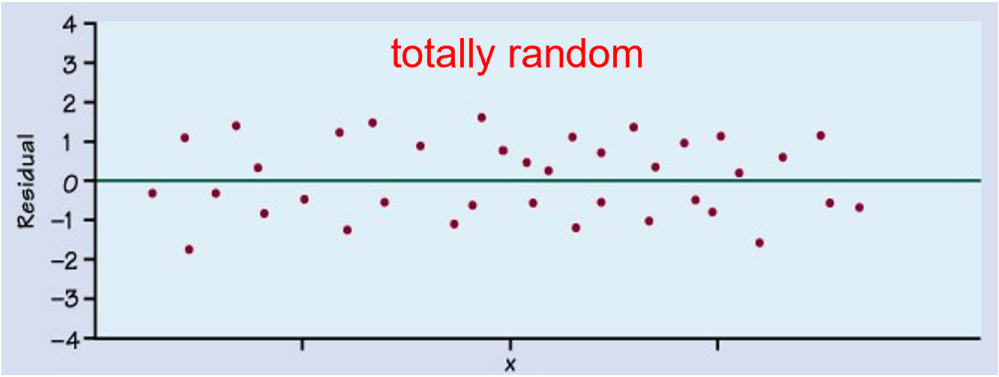
Scatterplot has high r^2 value, strong-looking regression line



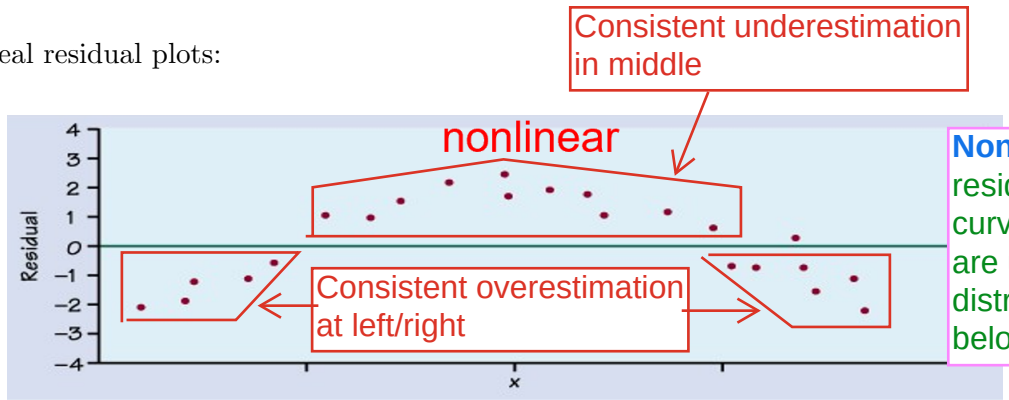
Residuals wide on left/right: regression line is only predicting well in the MIDDLE, not on sides

Ideal residual plot:

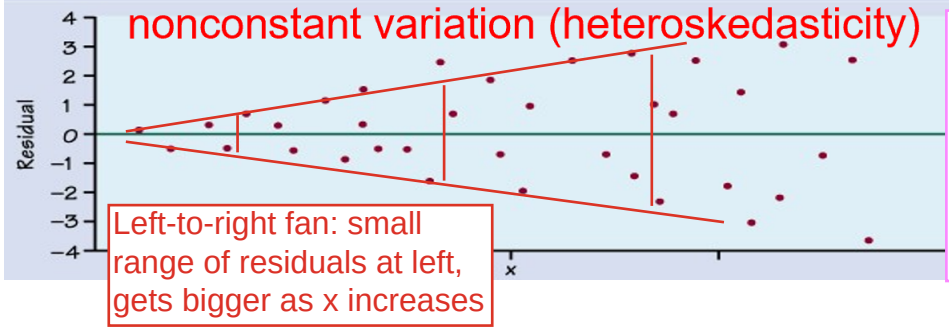
Residuals approximately form a rectangle, with the zero-line horizontal in the middle



Not-so-ideal residual plots:

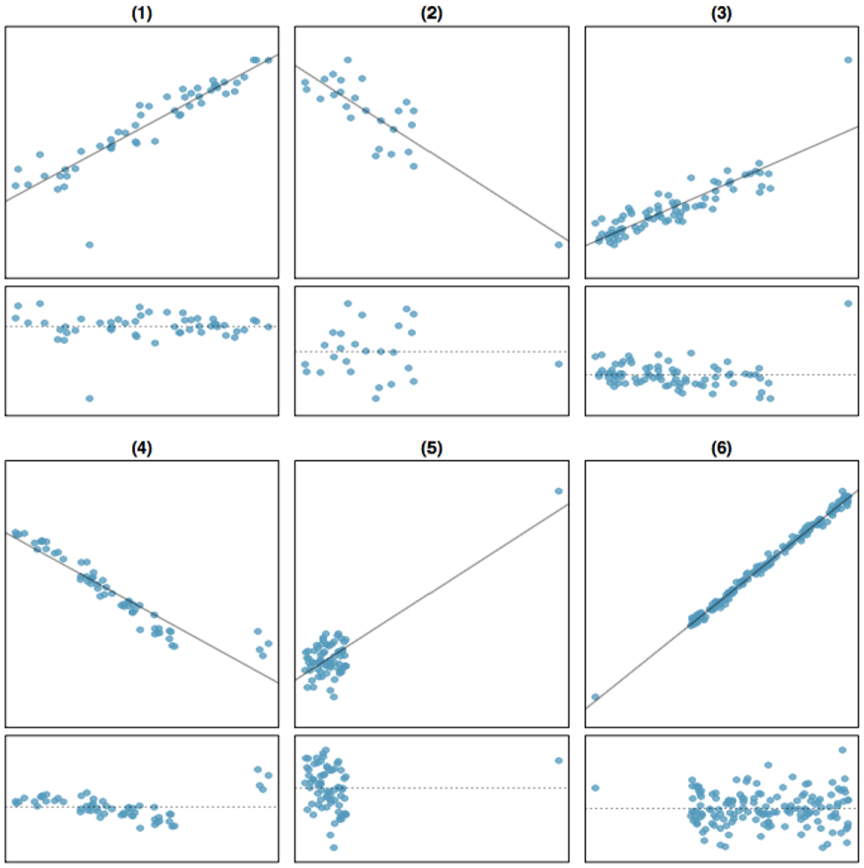


Nonlinear: residuals form a curved pattern, OR are unevenly distributed above + below the zero line



Nonconstant variance: residuals form 1 or more "fan shapes", i.e. "height" of residuals spread changes across x-values

Influence of outliers:



Residual plots can also help visualize outliers, even if they appear to follow the line
Keep in mind: regression line is likely heavily influenced by outliers! It's important to identify them for further investigation

Importance of Plotting Data

Data Set A

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Data Set B

x	10	8	13	9	11	14	6	4	12	7	5
y	9.14	8.14	8.74	8.77	9.26	8.10	6.13	3.10	9.13	7.26	4.74

Data Set C

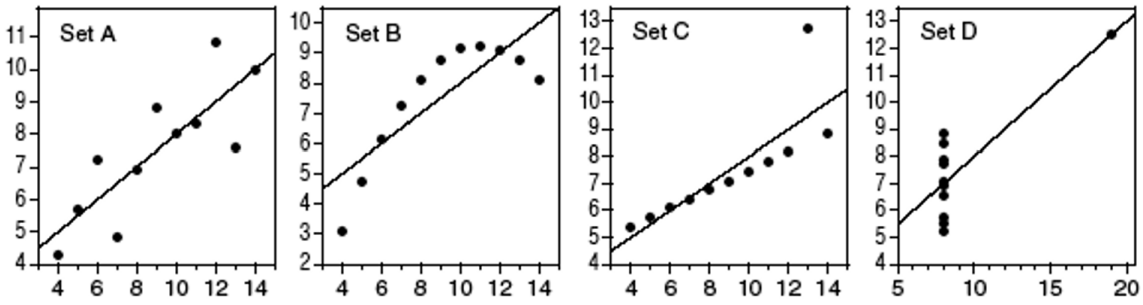
x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Data Set D

x	8	8	8	8	8	8	8	8	8	8	19
y	6.58	5.76	7.71	8.84	8.47	7.04	5.25	5.56	7.91	6.89	12.50

Source: Frank J. Anscombe, "Graphs in statistical analysis," The American Statistician, 27 (1973), pp. 17–21.

All have correlation $r = .816$, regression equation $\hat{y} = 3 + 0.5x$



Moral: Regression equation + correlation are NOT ENOUGH to fully describe a dataset!
(In other words, ALWAYS PLOT YOUR DATA)