ECOR 1010
Lectures 14 &15

Correlation and Regression

In the previous lectures:

- We only measured one variable for each sample, (x_i)
  - we just measured the weight of students with the N5/N20 machines, nothing else, we did not at the same time measure hair length, or height, etc.
- When we measure only one variable we use univariate statistics.

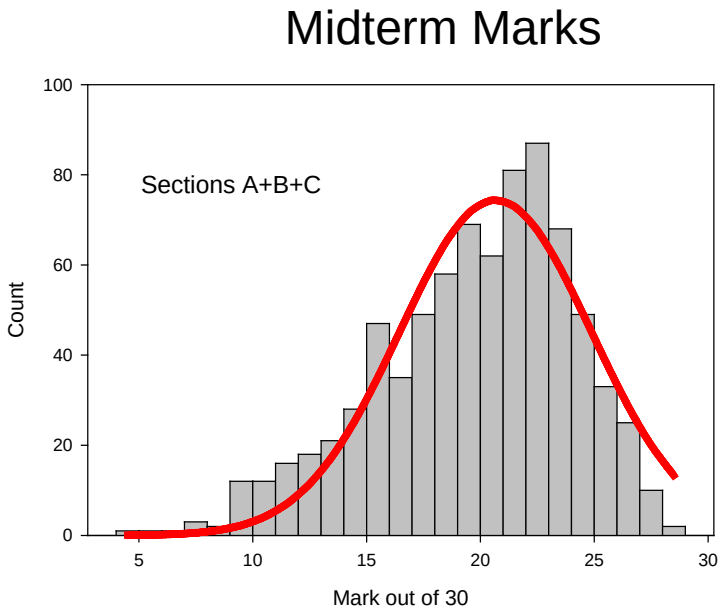
Bivariate relationships

- Now we look at the statistics we can use when two variables are measured
  - These are called bivariate statistics
- Bivariate observations come in pairs: (x_i, y_i)

- We want to know if there is a relationship between the variables x and y.
- We want to know if we can predict one from the other.

As an Example:

- We saw before that there was variability in the marks on the midterm
  - There was a spread in the marks
- You can imagine that there is also variability in class attendance
  - Not everyone comes to every class, and some miss more classes than others
- Can we relate marks on the exam to class attendance?

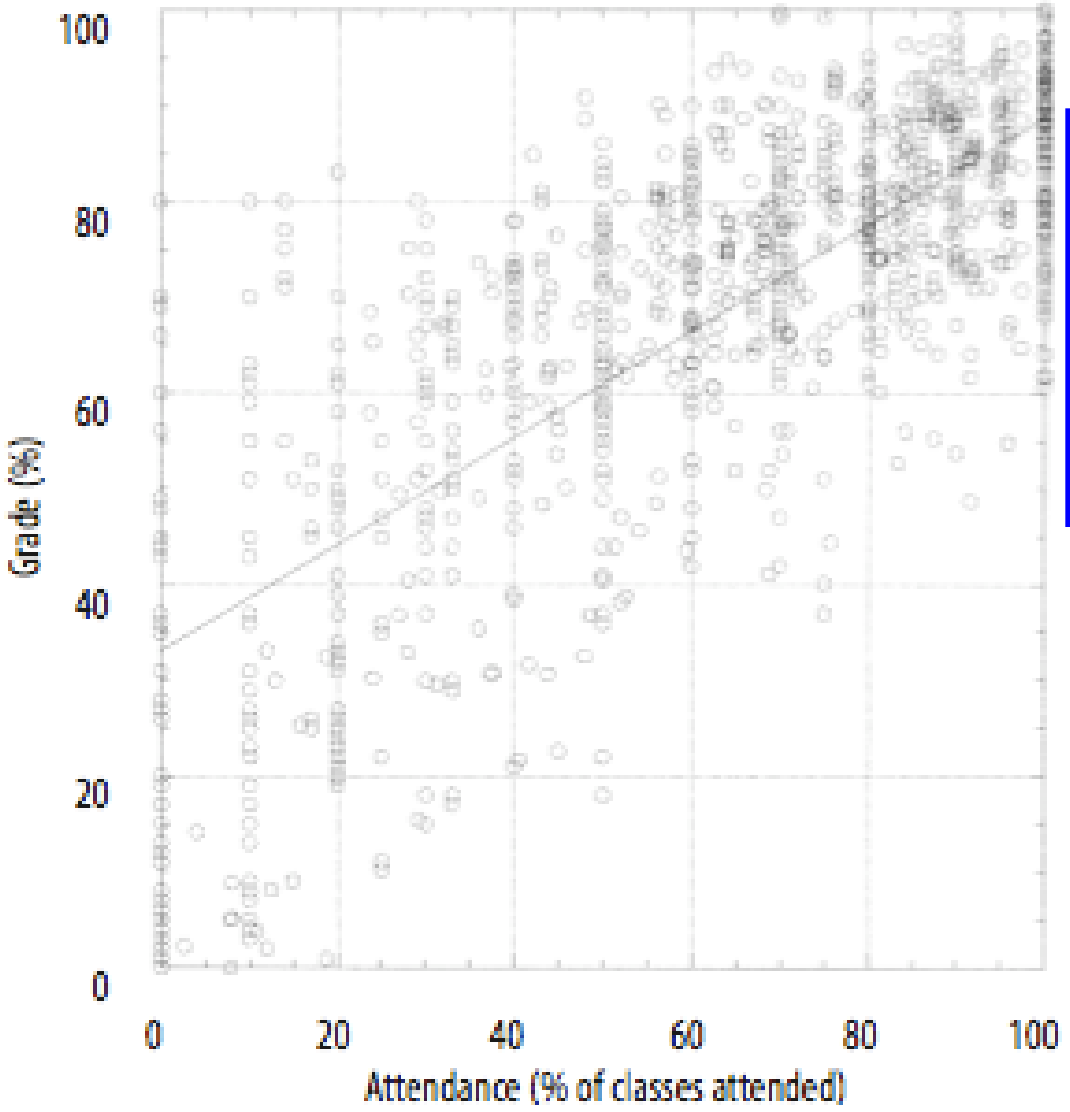


Of course, there are many reasons why people get the marks they do, but can we estimate the importance of attending class?

Showing Up: The Importance of Class Attendance for Academic Success in Introductory Science Courses

Randy Moore, Murray Jensen, Jay Hatch, Irene Duranczyk, Susan Staats, and Laura Koch
General College, University of Minnesota, Minneapolis, MN

The relation of class attendance and course grades in our Introductory Science classes. The size of this sample exceeded 1400. The equation for these data is $y = 33.1 + 0.55x$, and the correlation coefficient (r) = 0.78.



There is lots of variability (scatter), but, <u>on average</u>, we expect better marks for students who come to class more frequently.

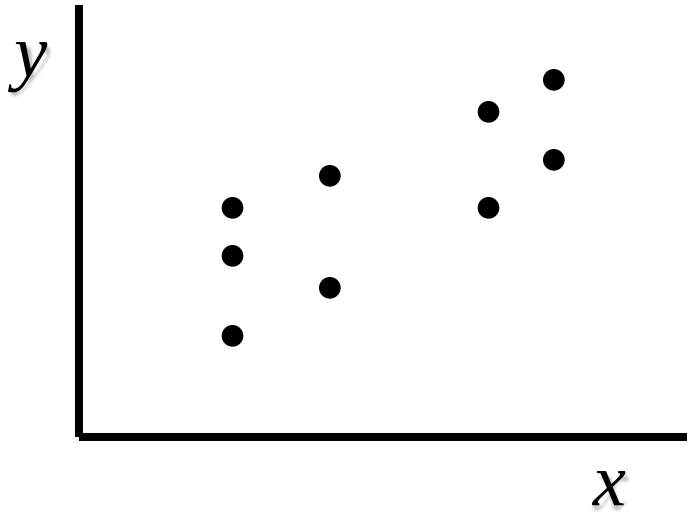
The words 'on average' are underlined in the box above because coming to more classes does not necessarily guarantee a higher mark for any individual – there are other things that we are not including in our analysis.

But, showing up explains 61% of the variability in the marks!

Bivariate relationships

- The distribution of bivariate measurements in a sample can be represented in a scatterplot (see below) with each point representing a pair of measurements (x, y) on some participant (i.e., the weight, and length of hair, of a student)
- There are 2 main questions that can be asked

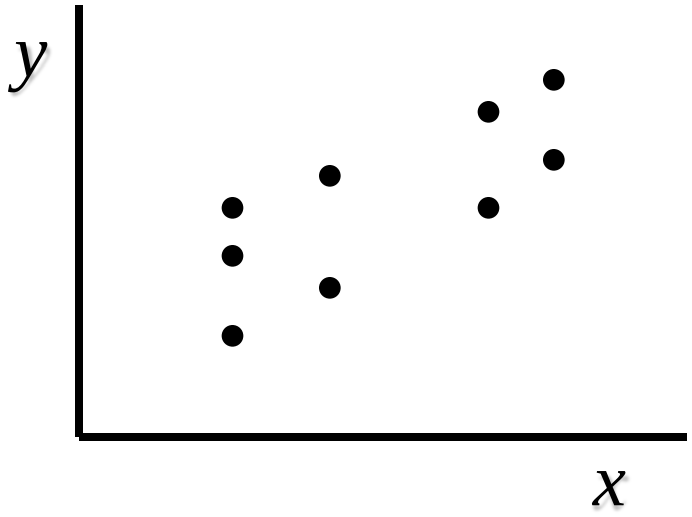
1) How strong is the relationship between the variables?
2) Can you predict the y's from the x's?



Bivariate relationships

- The answers to the first question are provided by the study of *correlation*
- The answers to the second question are provided by the study of *regression*

1) How strong is the relationship between the variables?
2) Can you predict the *y*'s from the *x*'s?



Linear Correlation

- Correlation is a measure of the **strength of a relationship**
  - How well does a “line of best fit” represent a trend in a data set?
  - How “correlated” are the two variables?

Linear Correlation

- How “good” is the assumption that a simple straight-line (i.e., linear) relationship exists between *x* and *y*?
- We can compute a correlation coefficient (*r*)

$$-1 \leq r \leq 1$$

  - This is a measure of how ‘connected’ *x* and *y* are in terms of the linear relationship assumption
  - Another way of saying it: *r* indicates how well a line of best fit correlates with the data

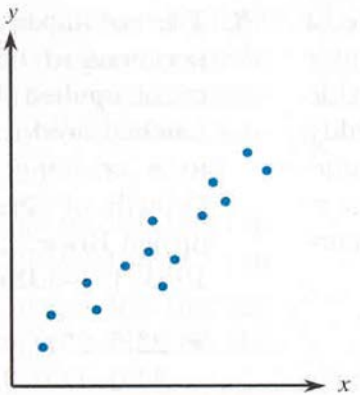
Some important values of the Correlation Coefficient: *r*

- When *r* is close to 1, it means that
  - *x* is large when *y* is large, and *x* is small when *y* is small
  - The plot of (*x*, *y*) is **tightly packed**
  - We say this is a 'Good' Positive Correlation
- When *r* is close to -1, it means that
  - *x* is large when *y* is small, and *x* is small when *y* is large
  - The plot of (*x*, *y*) is **tightly packed**
  - We say this is a 'Good' Negative Correlation
- If *r* is near zero
  - Little or no (linear) relationship exists between *x* and *y*

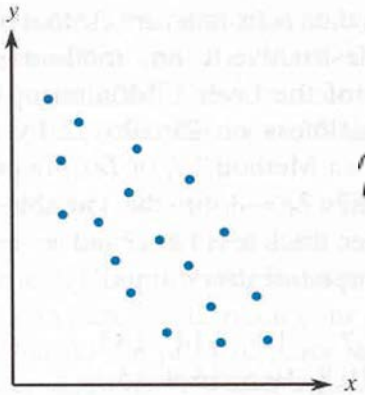
$$-1 \leq r \leq 1$$

Correlation Coefficient Examples

$r > 0$



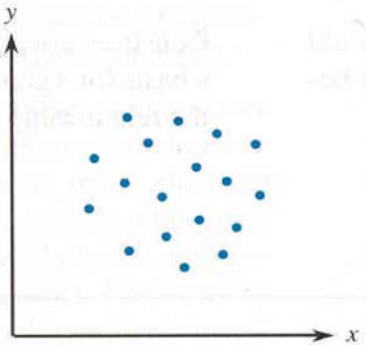
(a)



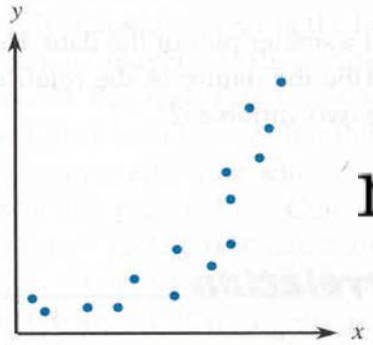
(b)

$r < 0$

$r \approx 0$



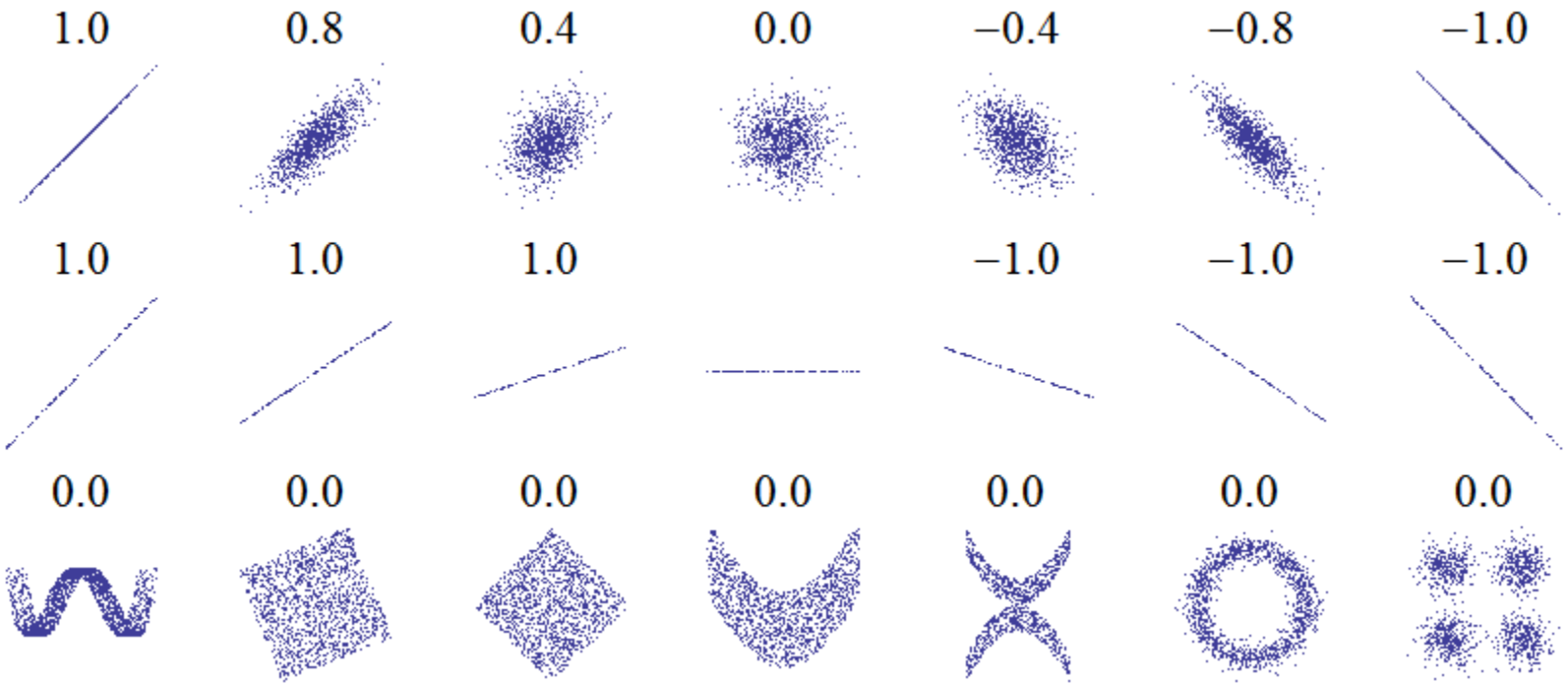
(c)



(d)

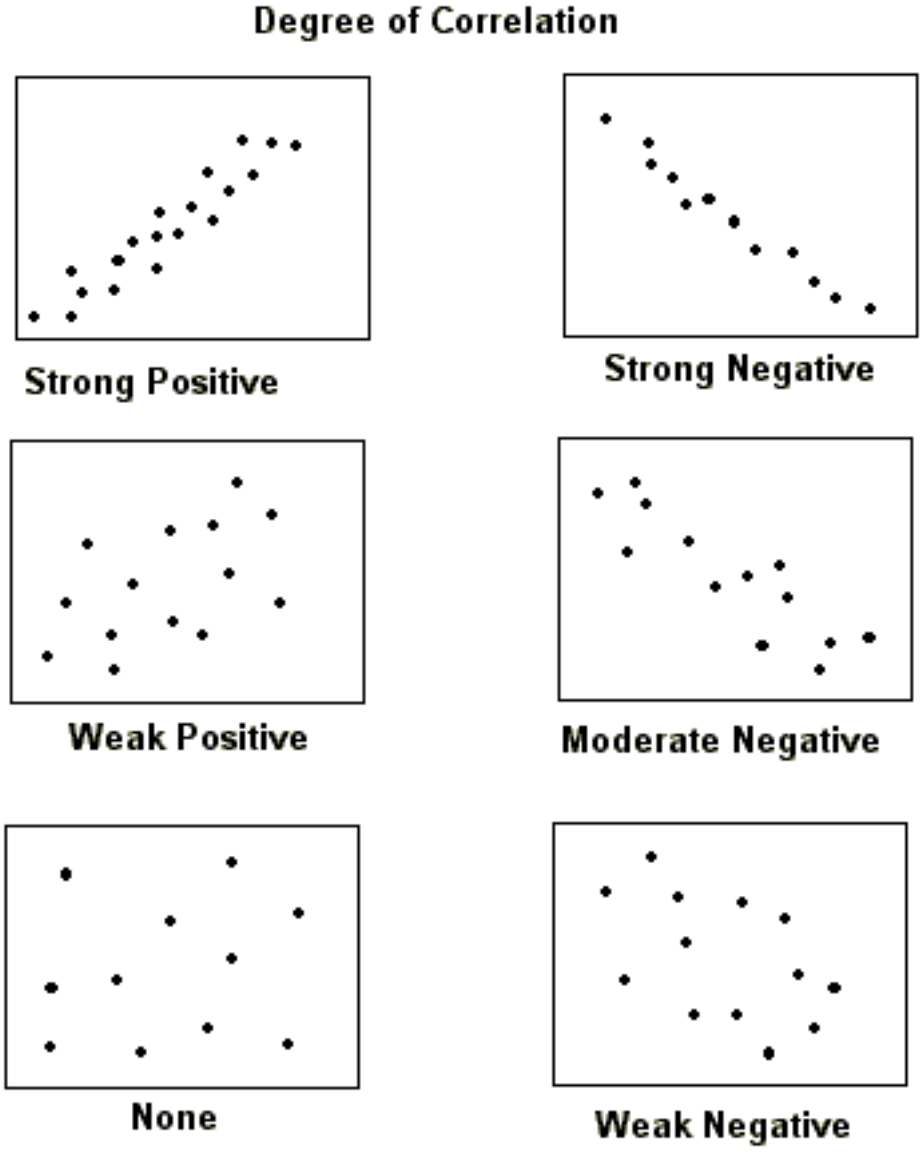
not linear

Correlation Coefficient Examples



Several sets of (*x*, *y*) points, with the Pearson correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero. (Wikipedia)

Correlation Coefficient Examples



Computing the Correlation Coefficient

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

NO UNITS !

n = number of data points

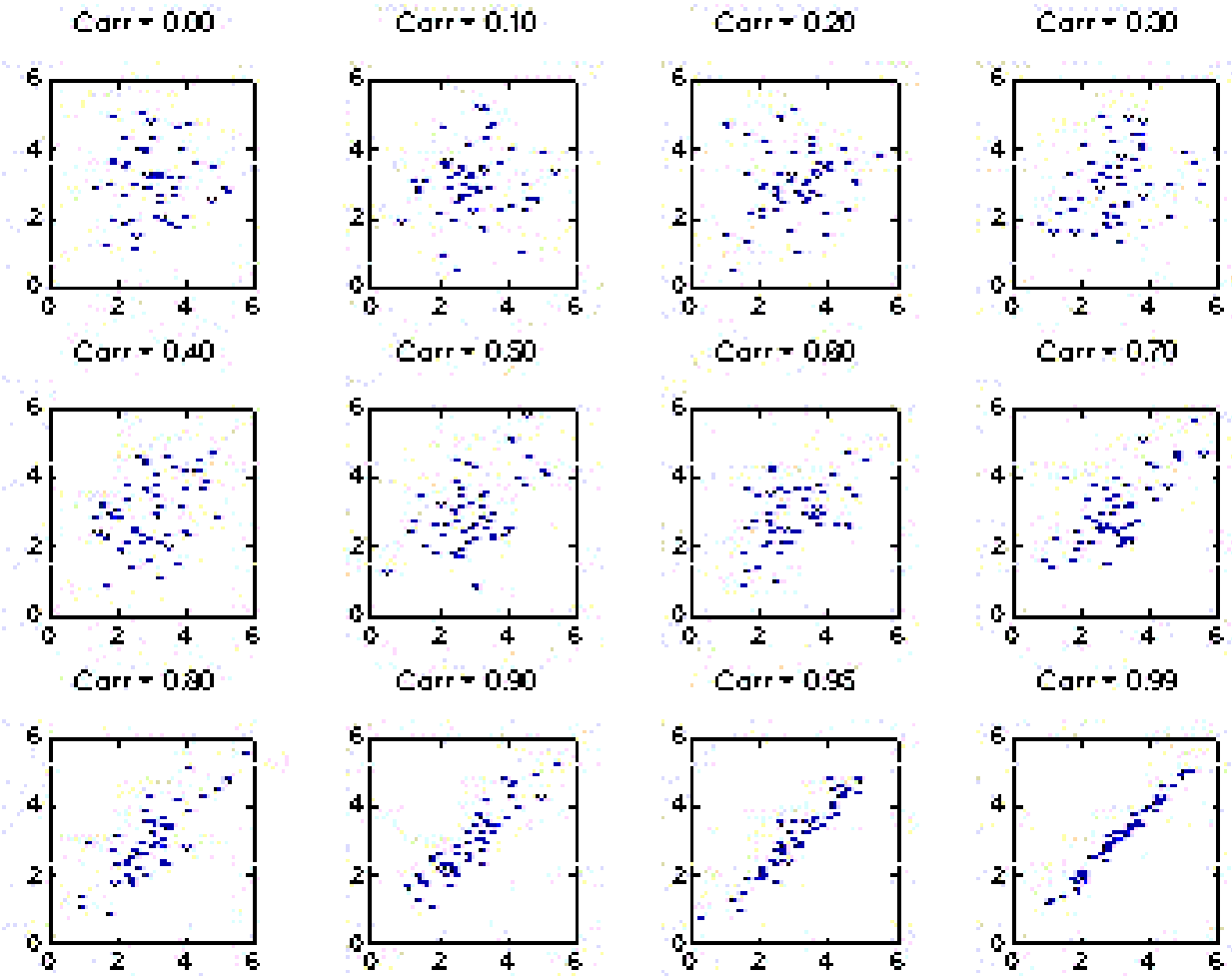
$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

s_x = standard deviation of x

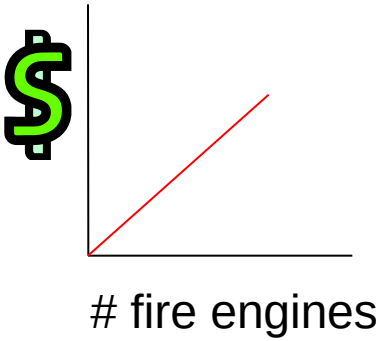
s_y = standard deviation of y

Correlation Coefficient Examples



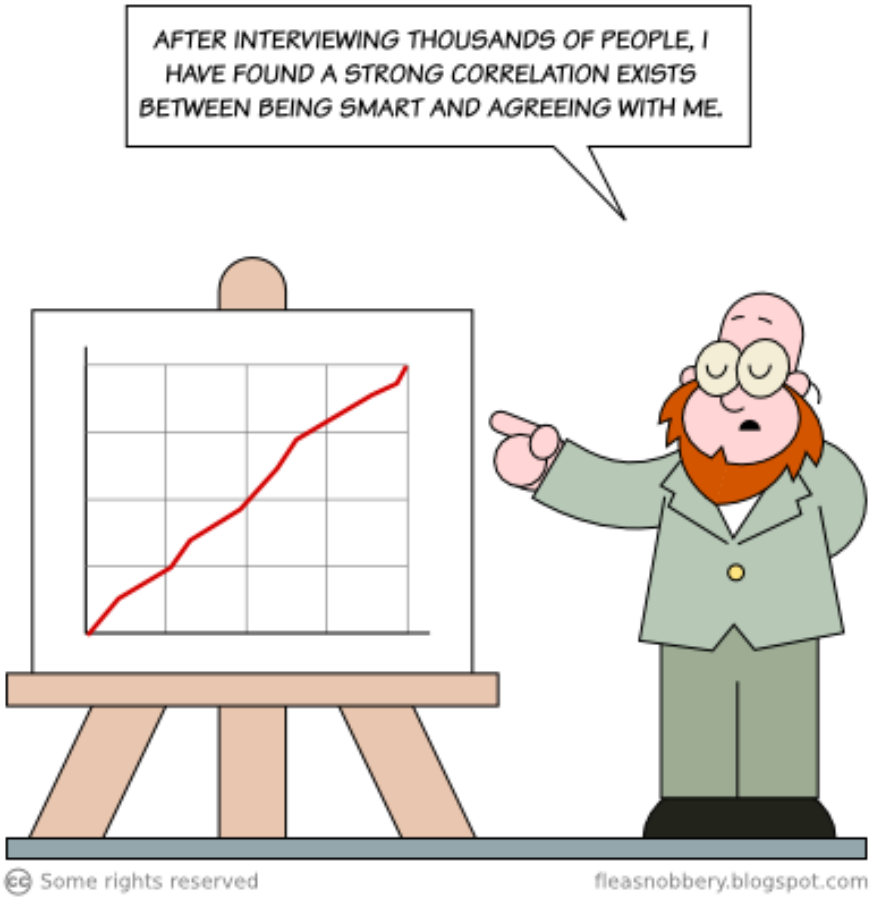
Correlation and Causality

- Get data on all the fires in Ottawa for the last ten years.
- Correlate the number of fire engines at each fire and the damages in dollars at each fire.
- Note the significant correlation between number of fire engines and the amount of damage.
- Conclude that fire engines cause the damage.



Correlation and Causality

- Just because two variables are correlated does not mean there is any connection between them



Regression

- The correlation coefficient expresses the strength of the relationship between *x* and *y* without worrying about units or directional relationships. It is the answer to the first question posed earlier: what is the strength of the relationship?
- However, if *x* and *y* are correlated, this implies that we can use information about *x* to *predict* the values of *y*
- If we want to *predict y* based on *x* we need more than r_{xy}…we need to perform a *regression*

Regression

Regression is a process by which a mathematical model is fitted to a set of data

- to **test** whether the model provides a reasonable description of the observations, and
- to give us confidence that **predictions** we make with the model will be good.

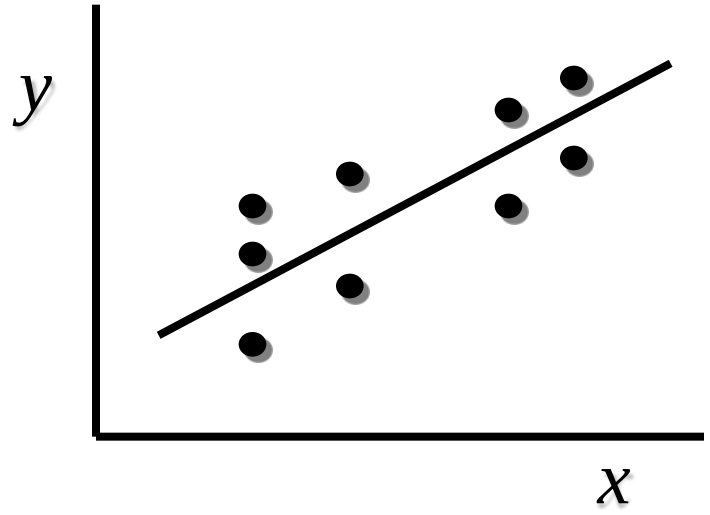
Father of Regression

- Sir Francis Galton (cousin of Charles Darwin) developed techniques for fingerprinting, weather forecasting, as well as regression and correlation
  - 1822-1911
- Developed regression by performing an experiment using five groups of sweet peas
  - Noticed that the average weight of sweet pea seedlings "regressed" towards the average of all of the peas, and not the parental group from which they came
  - Applied a similar technique to predict human heredity

→ "Regression to the Mean"

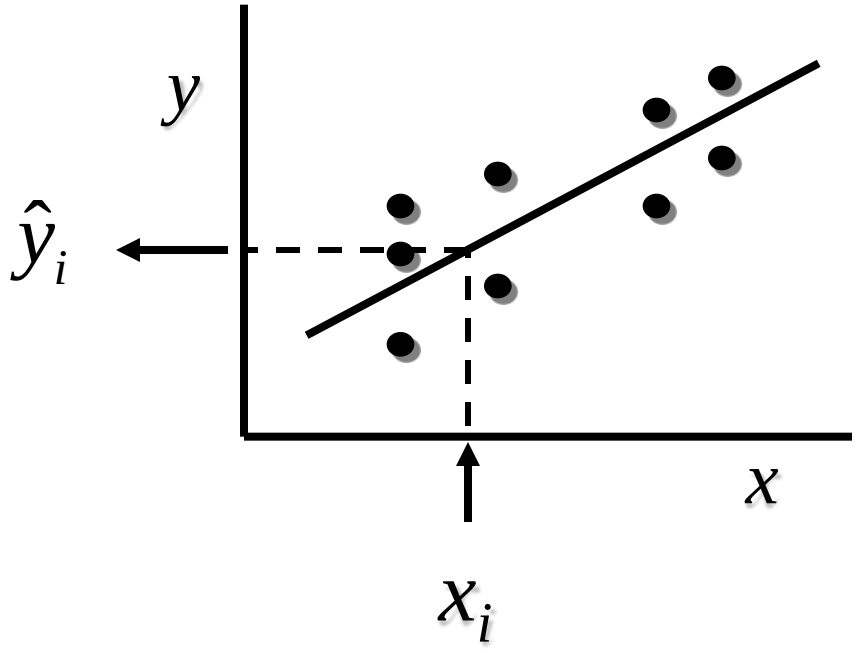
Regression

- Suppose we want to find the straight line relationship that best predicts the values of *y*, given *x*…this is called the *regression line*



Regression

- Once we figure out what the line should be, for any given *x* we can find a predicted *y*
- The predicted value is called $\hat{y}$



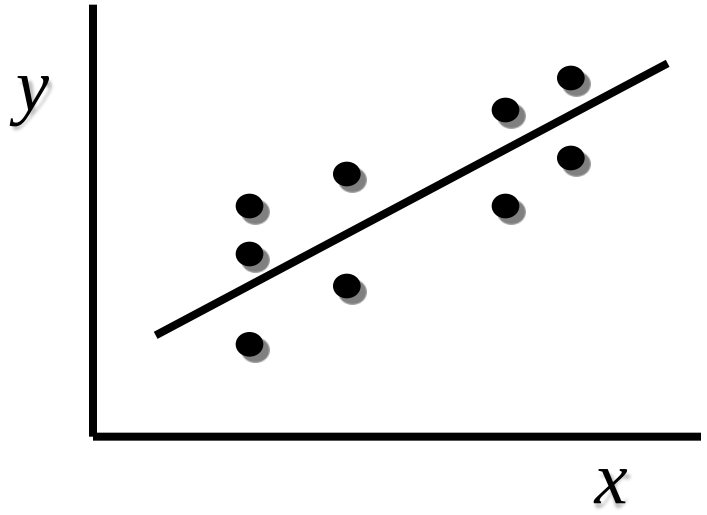
Regression

- The straight line is one we have made up to express the relationship between *x* and *y* that we see in the data
- The line has an equation

$$\hat{y} = mx + b$$

slope

intercept



- Find a “line of best fit”
  - Use a parameterized mathematical model

$$\hat{y} = mx + b$$

where

$\hat{y} =$ predicted dependent variable

$x =$ independent variable

$m =$ slope

$b =$ y intercept

Regression

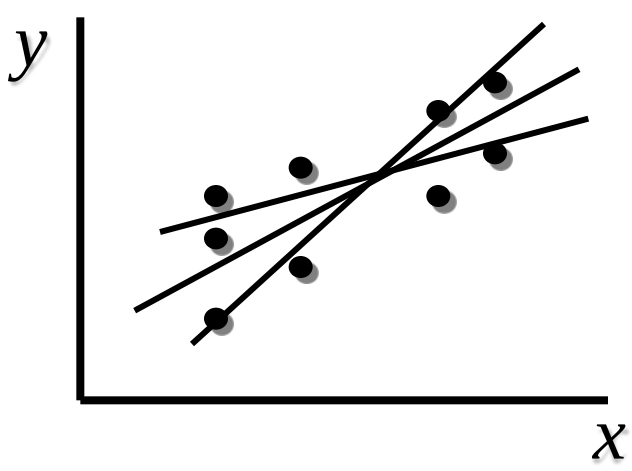
- The slope, *m*, and intercept, *b*, of the regression line each have their own interpretation

$$\hat{y} = mx + b$$

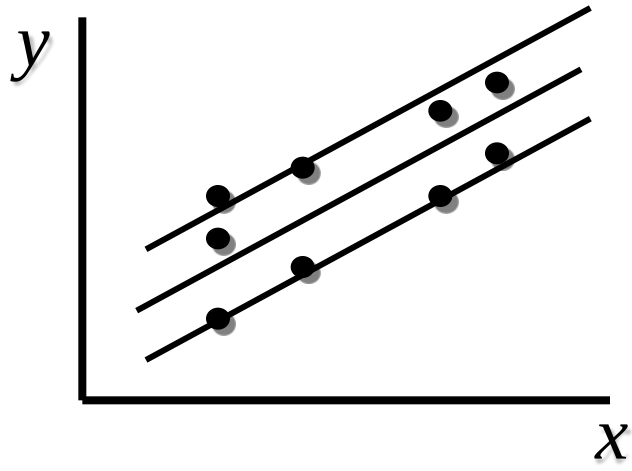
- *intercept*: *b* = value of $\hat{y}$ when *x* = 0
- *slope*: *m* = difference in $\hat{y}$ associated with a change of *x* by 1 unit

Regression

- Each different combination of *m* and *b* create a different regression line



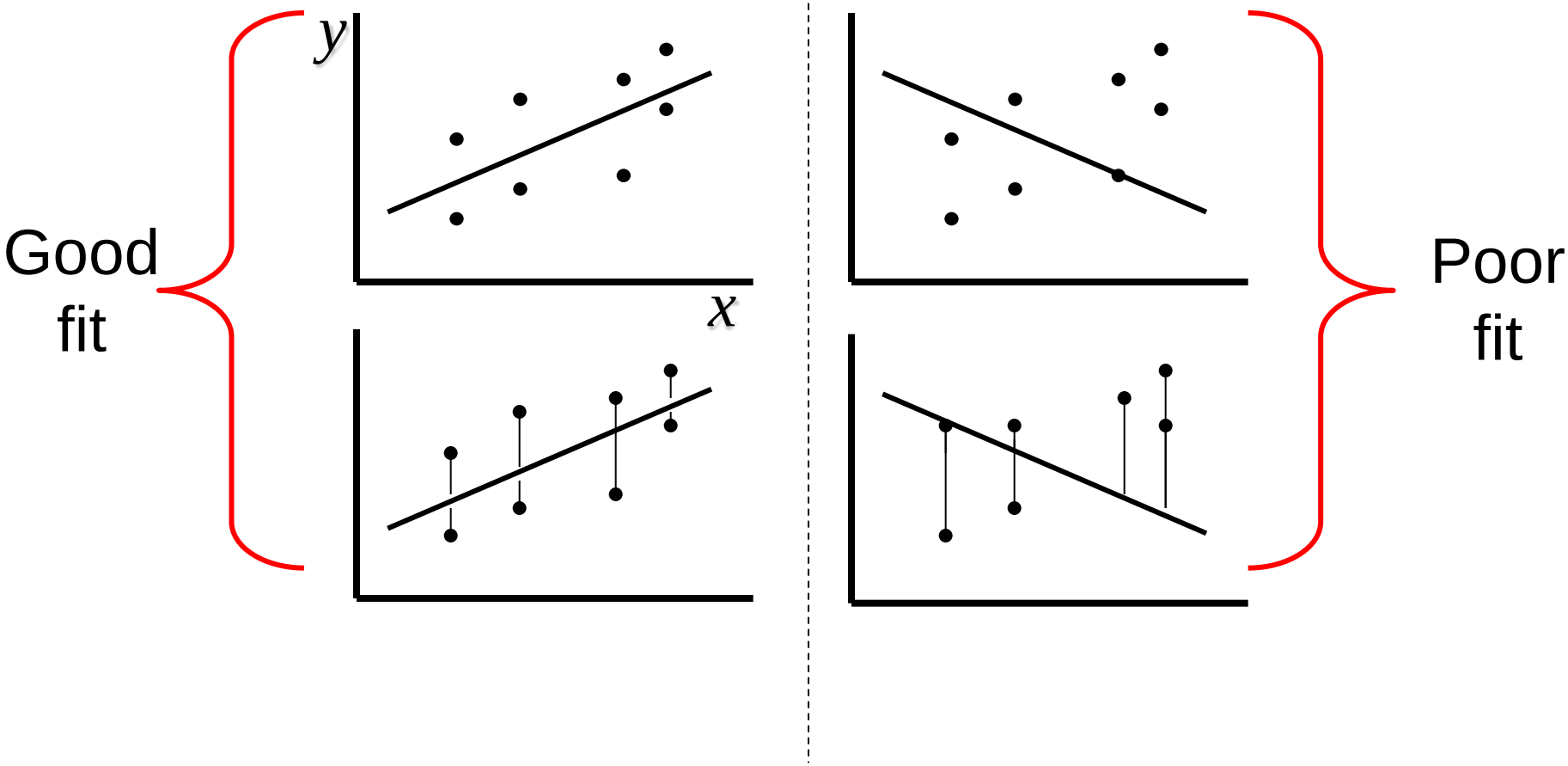
Different *m*'s



Different *b*'s

Regression

- The *best* regression line is the one that minimizes the *sum of squared deviations* of the data from the line



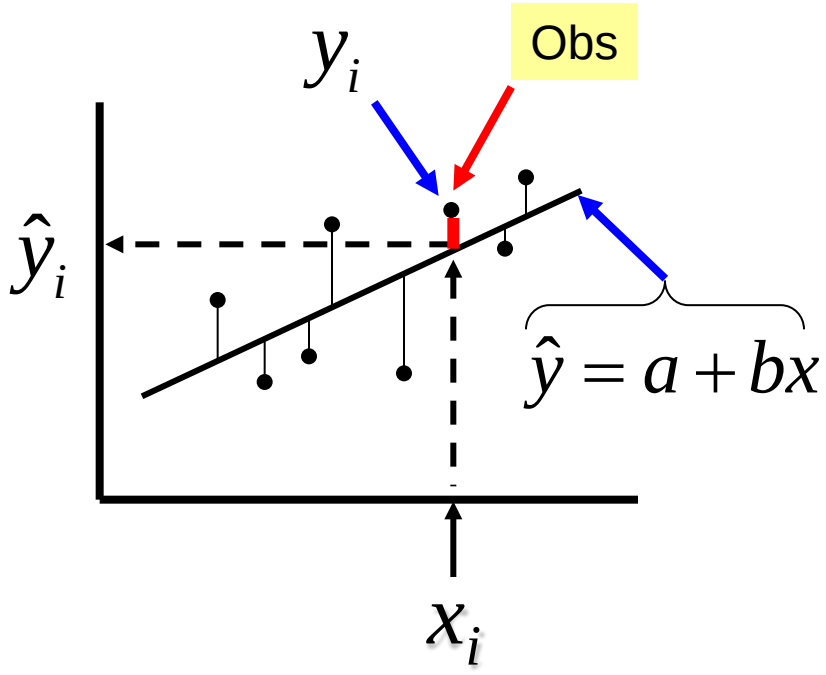
Regression

- For each datapoint connected with an observation x_i, the appropriate "deviation" is the deviation of the observed y_i from the value, $\hat{y}_i$, calculated from the regression equation

- The deviations are, therefore, given by

$$e_i = y_i - \hat{y}_i$$

and are called the "residuals" or "errors" of the fitted line for the data



Linear Regression

- The error, or **residual**, is the difference between the observed point, y, and the calculated point on the line, $\hat{y}$, at each x

$$e_i = y_i - \hat{y}_i$$ = Obs - Calc = Measured - Predicted

where

$\hat{y}_i$ = value of y predicted by the model

y_i = the i^{th} measured data point

What is the "Best Fit"?

- Consider the following **S**um of **S**quared **E**rrors

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

Obs - Calc

- SSE is a measure of how well the model fits the data
- The parameters *m* and *b* of the best fitting line for a particular sample are determined by finding the line that *minimizes* SSE
- Called **least-squares fit/regression**

Regression

- Try some of the Java applets concerning regression at http://onlinestatbook.com/rvls/index.html

- The link called “Regression by Eye” shows how different regression lines have different associated SSE’s (In the applet instead of SSE, they use MSE, which stands for mean square error. This is the sum of the squared errors, SSE, divided by n-1.)

Least-squares Fit/Regression

(without proof)

For $\hat{y} = mx + b$,

The 'least-squares' slope

and intercept are given by:

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

and $b = \overline{y} - m\overline{x}$

n = number of data points

Eg. Fuel Consumption and Car Weight: Is there a relationship? Can we predict one from the other?

Model	Weight (hundreds of Lbf)	City Mileage (mpg)
Ford Festiva	18	31
Honda Civic	20	36
Toyota Matrix	27	30
Acura RSX	27	27
Mazda Protégé 5	27	29
Subaru Impreza WRX	31	24
Pontiac Firebird	32	21
BMW 330 Ci	33	21
Linxoln Copntinental	39	14
Chevrolet Suburban	51	11

Which is the dependent variable?

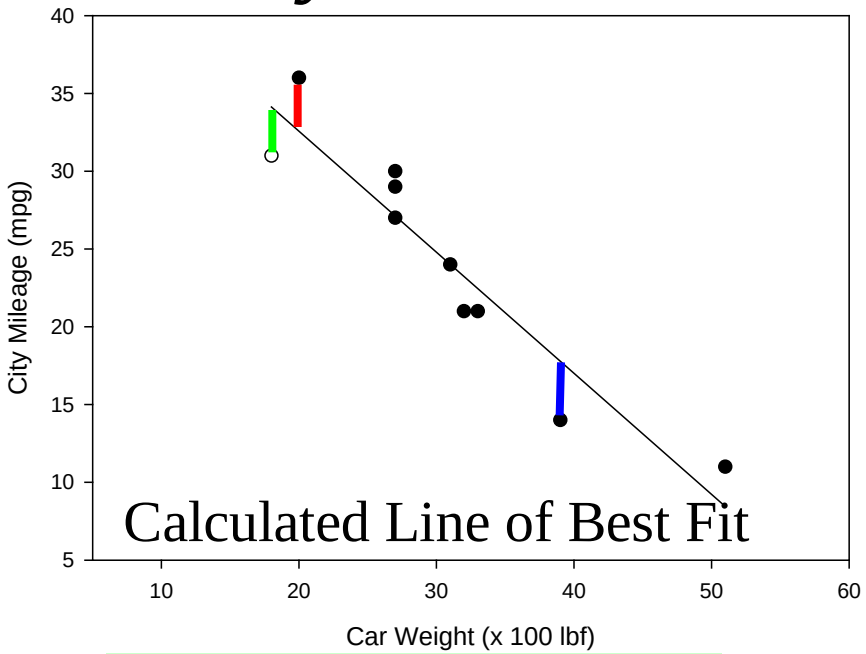
Which should be plotted on the y axis?

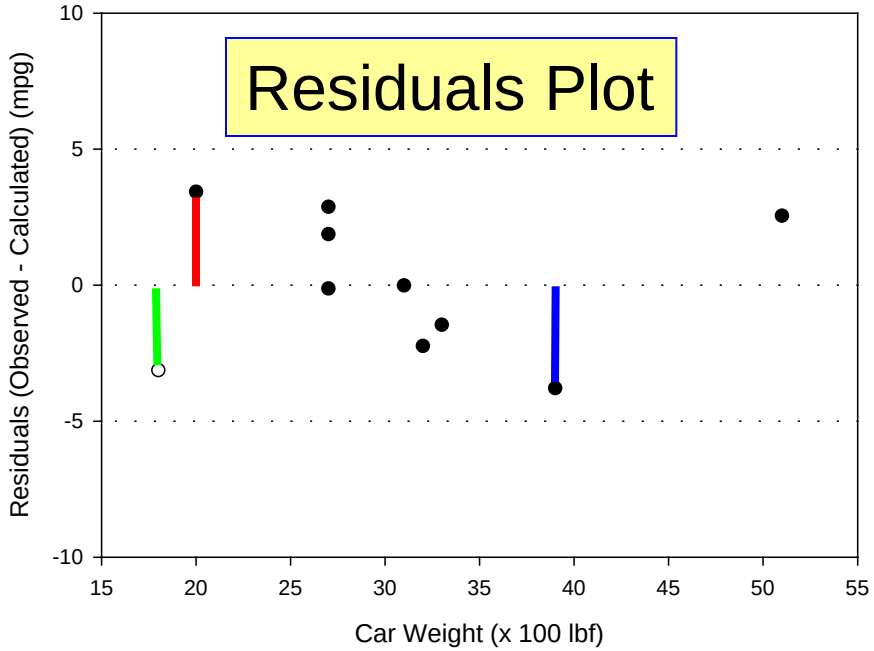
Line of Best Fit

Obs - Calc

$\hat{y} = mx + b$

$(y_i - \hat{y}_i)$





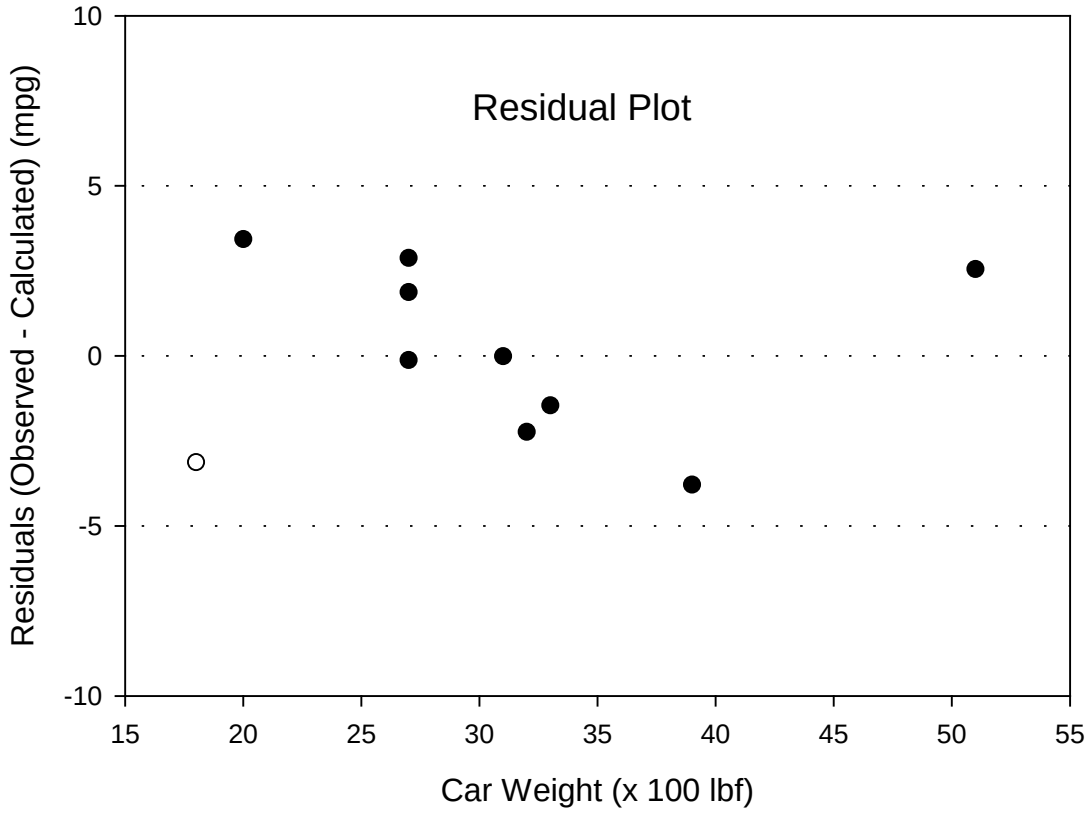
The line through the data is the "calculated data".

Residuals = Observed Values – Calculated Values

Least-Squares Regression minimizes SSE = $\sum (y - \hat{y})^2$

Obs - Calc

Residuals

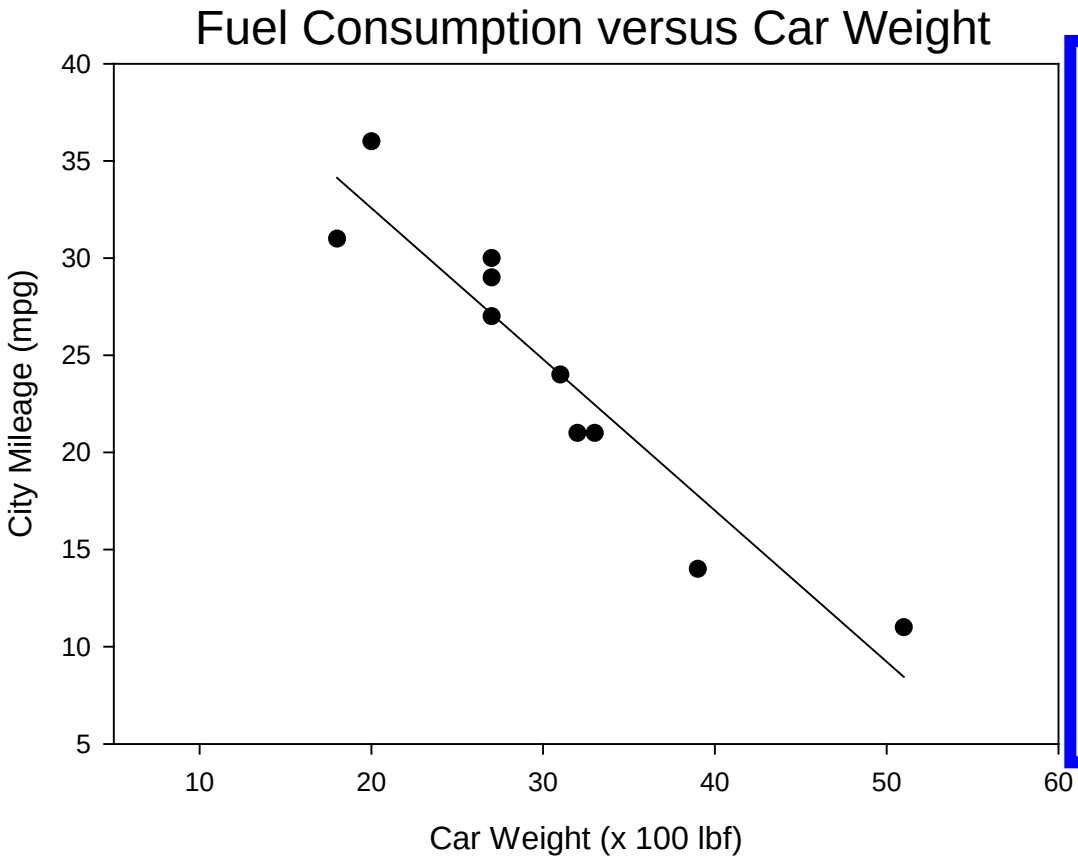


The Residuals are equally (normally) distributed about 0, which is what we want to see.

Coefficient of Determination

- Often use r^2 to estimate the quality of a fit
  - Represents the proportion of the variation of the observed *y* data that can be explained by the model (discussed later)
  - Lies in the range [0,1]
  - Close to 1 means the assumed function (i.e., straight line in this example) is a good fit to the observed data

Coefficient of Determination



- The cars in the sample have Mileage values that vary from 11 to 36: this is the variability in the *y*-values.
- 89% of the variability in the *y* data can be explained by the straight-line relationship (why we can say this is explained on the next slides)

$$\hat{y} = -(0.8 \pm 0.1)x + (48 \pm 3) \quad [\text{mpg}]$$

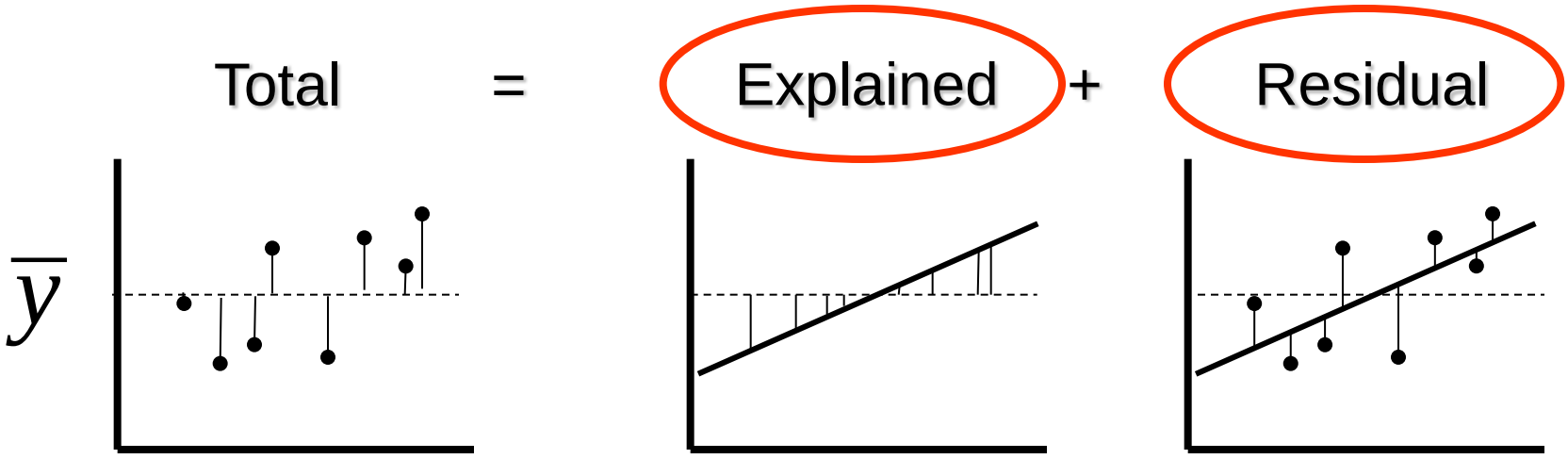
$$r^2 = 0.89$$

Relationship of the 'Coefficient of Determination' to Explained Variation

- Only part of the variation in a measurement (data point) can be explained by regression.
- There are two sources of variation:
  - Variation explained by the regression model
  - Variation not explained by the regression model
- Total variation = Explained + Unexplained

Accuracy of Prediction

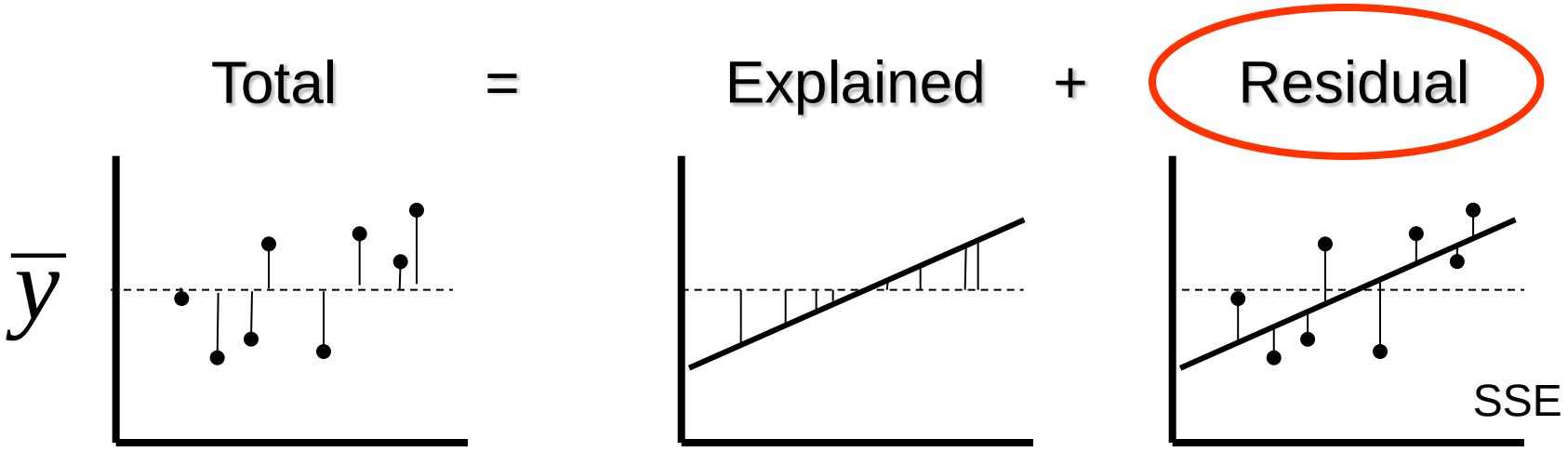
- Here is what is going on… there are 2 components of variation that together make up the total variation in *y*



Variation is measured relative to the mean value of all of the y measurements: $\overline{y}$

Accuracy of Prediction

- You have seen the “residual” variation before, it is the variation left over after fitting the regression line to the data



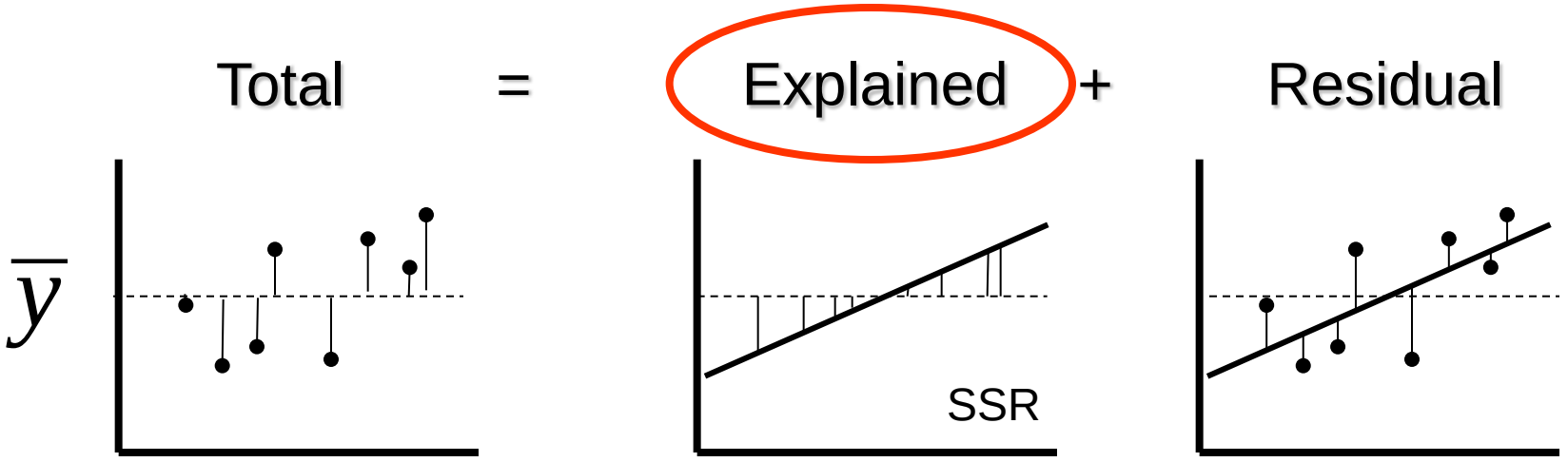
- Residual variation is quantified by the SSE and is minimized in a regression

SSE = sum of squared errors (residuals), or what is left over (resides) that has not been explained after you factor out the linear part.

$$\sum (y - \hat{y})^2$$

Accuracy of Prediction

- "Explained" variation is the variation in y values that is explained by the regression line



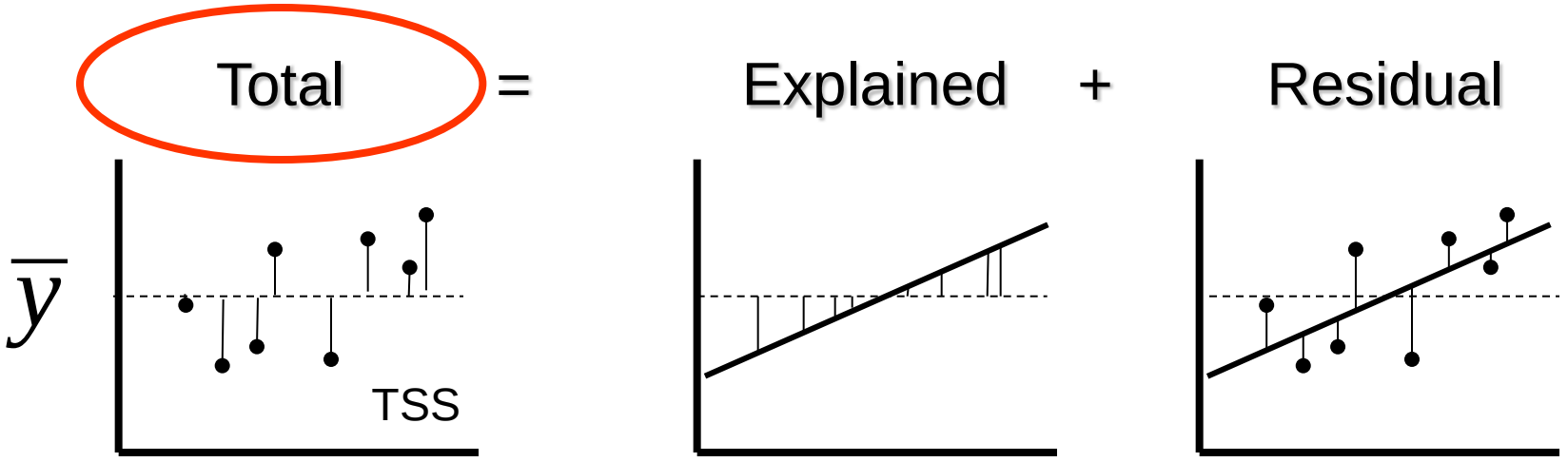
- Explained variation is quantified by SSR

SSR = sum of squares of the deviations from the regression line to the mean of all the y`s.

$$\sum (\hat{y} - \overline{y})^2$$

Accuracy of Prediction

- "Total" variation is the overall variation in y values without regard to any information about the *x* values



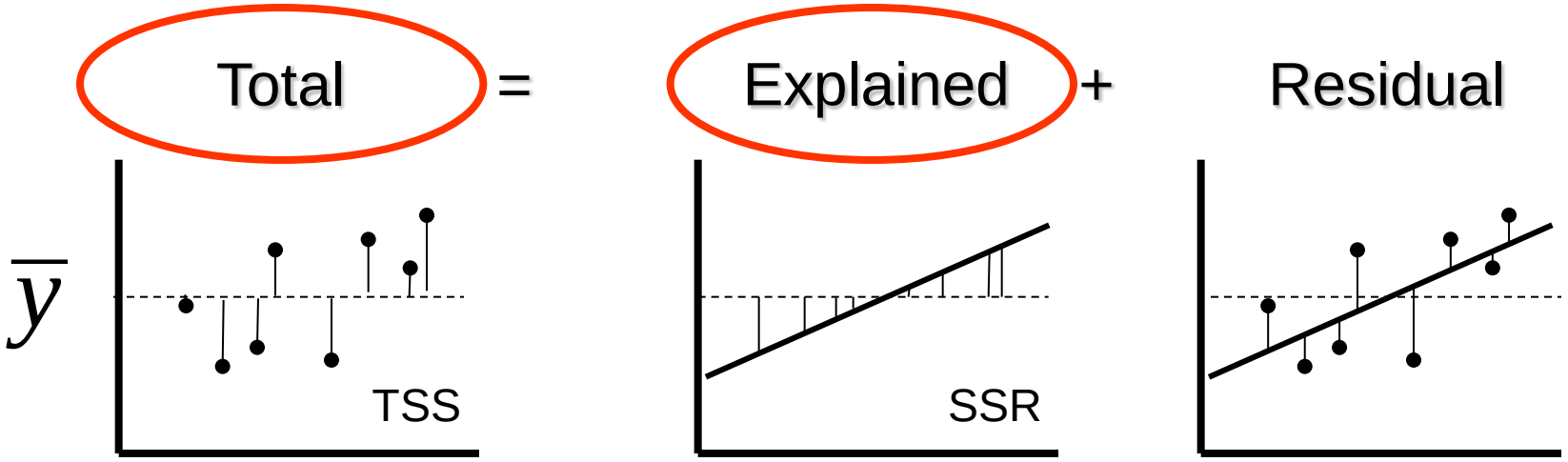
- The total variation is quantified by TSS

TSS = total sum of squares from the mean of all of the y data

$$\sum (y - \overline{y})^2$$

Accuracy of Prediction

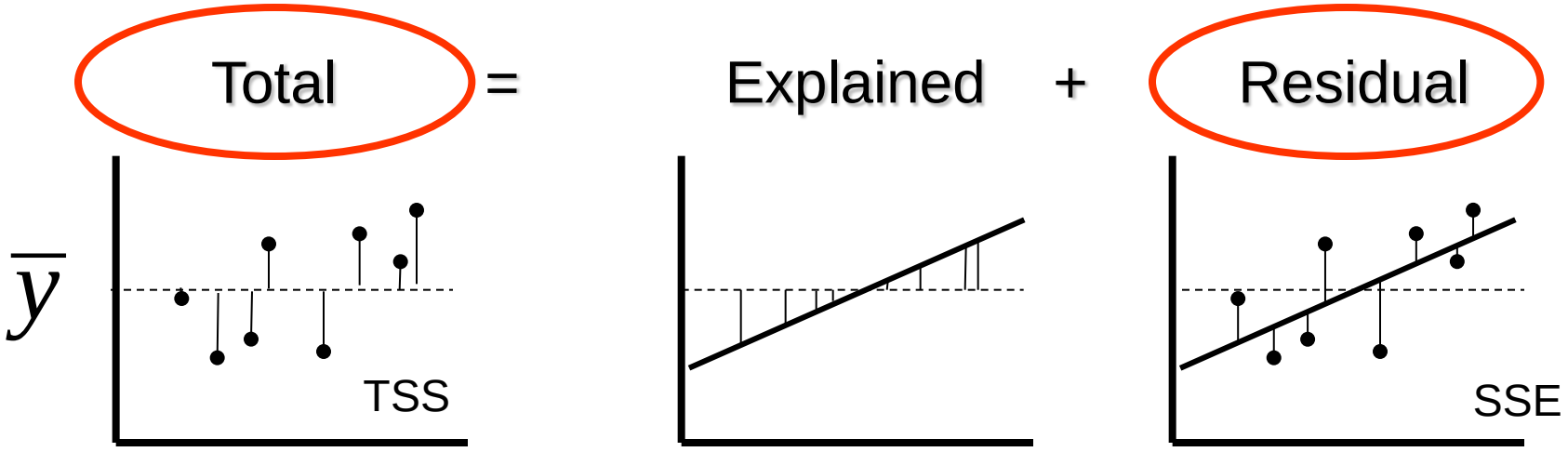
- So the *ratio* of explained variation to total variation characterizes how well the regression describes the data



- Explained variation $\frac{SSR}{TSS}$

Accuracy of Prediction

- Likewise, we get the *proportion of unexplained variation* by dividing SSE by TSS



- Unexplained variation = $\frac{\text{SSE}}{\text{TSS}}$

Relationship to Variation

$$\sum (y-\bar{y})^2 = \sum (\hat{y}-\bar{y})^2 + \sum (y-\hat{y})^2$$

- TSS = SSR + SSE
- TSS = total sum of squares
- SSR = sum of squares from regression
- SSE = sum of squared errors (residuals), or what is left over (the residue) after you factor out the part explained by the assumed model.
  - In this case, the assumed model is a straight line.

Accuracy of Prediction

- It should not be surprising that the proportion of explained variation plus the proportion of unexplained variation equals 1
- So the proportions of explained and unexplained variation work together to account for all the variation in *y*…
- You can see this in the simulation, available at the RVLS website:

http://onlinestatbook.com/rvls/index.html

- called the "Components of *r* demonstration"…go there and try it out

Accuracy of Prediction

- The *accuracy of prediction* for a regression is quantified by the proportion of explained variation

$$\frac{SSR}{TSS}$$

- Think of this as the proportion the total variation in *y* values that is explained by the variation in the *x* values…the explanation is due to the correlation between *x* and *y*

Accuracy of Prediction

- There is one last amazing result concerning the proportion of variation explained by a regression…

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\text{SSR}}{\text{TSS}}$$

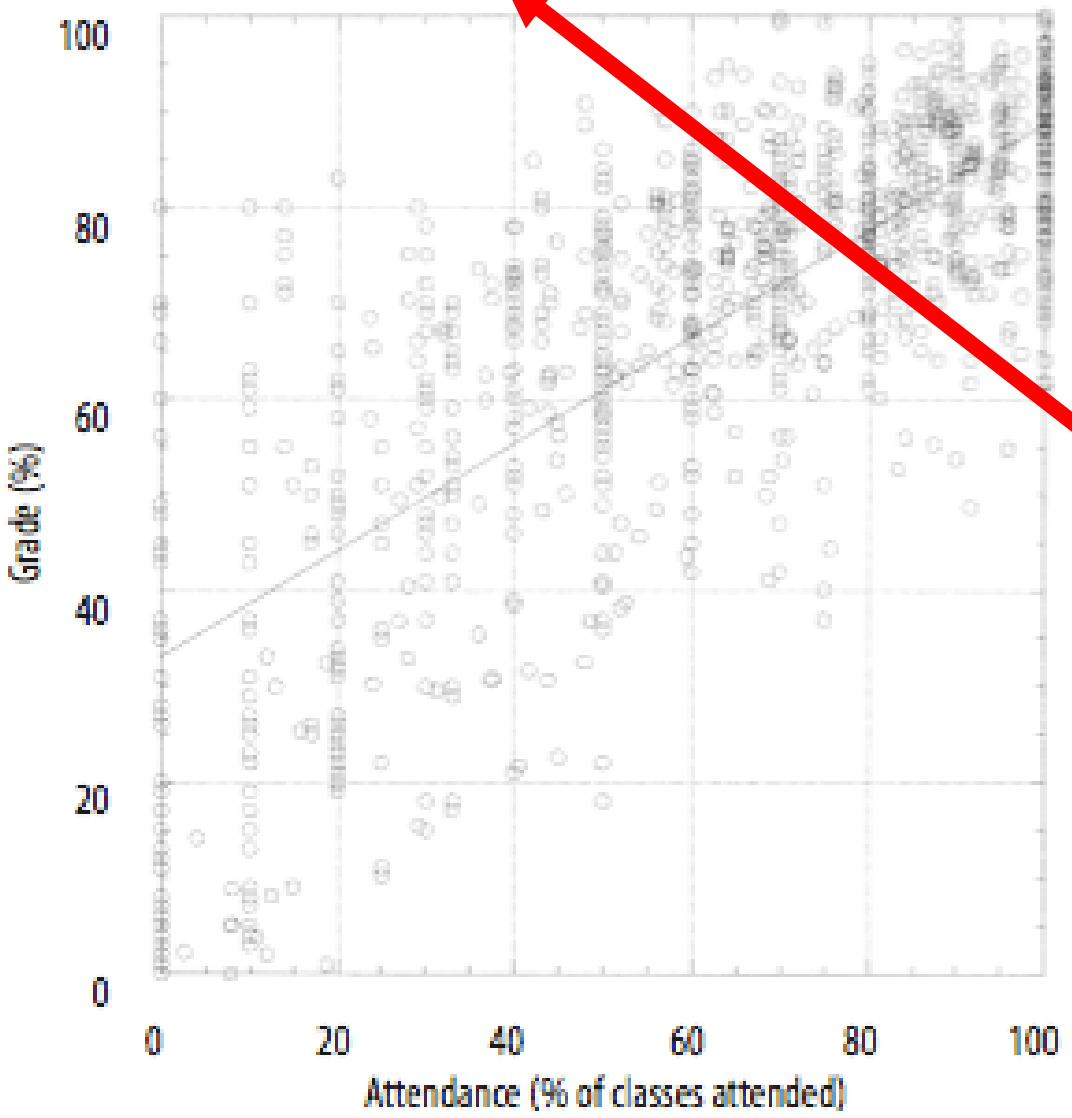
- In other words, the proportion of explained variation for the straight-line fit equals the square of the linear correlation coefficient!
- This ties together correlation and regression

r^2 is called the Coefficient of Determination

Showing Up: The Importance of Class Attendance for Academic Success in Introductory Science Courses

Randy Moore, Murray Jensen, Jay Hatch, Irene Duranczyk, Susan Staats, and Laura Koch
General College, University of Minnesota, Minneapolis, MN

The relation of class attendance and course grades in our Introductory Science classes. The size of this sample exceeded 1400. The equation for these data is $y = 33.1 + 0.55x$, and the correlation coefficient (r) = 0.78.

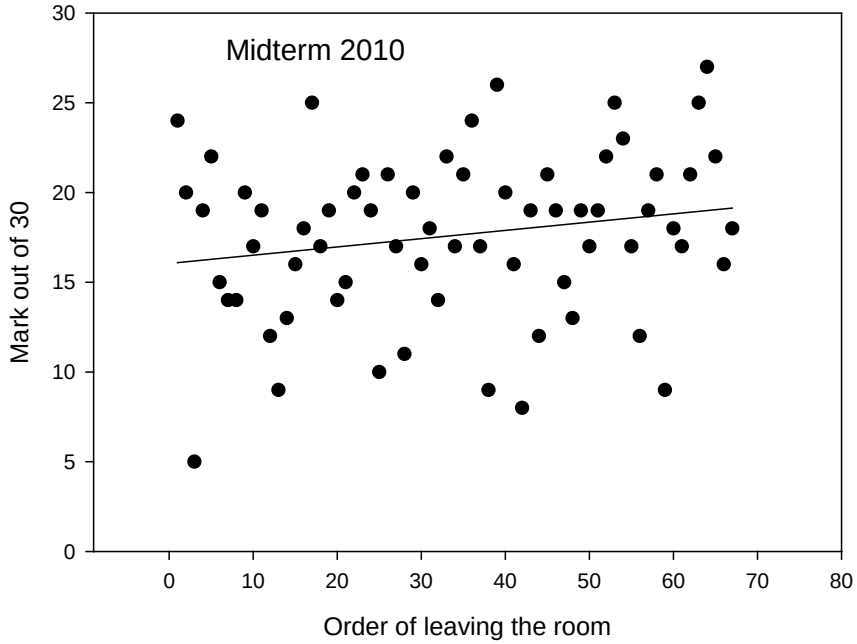


The proportion of the total variation in marks explained by the linear relationship with class attendance is $r^2=(0.78)^2$ or 61%.

Closer to home:

- Marks for the midterm versus the order that students left the room.

$$r^2 = 0.0375$$



- Roughly 4% of the variability in the marks can be explained by the order students left the room.
- An Aside: On average, students who stayed till the end of the exam got 3 more marks than those who left early.

Linear Least Squares Regressions/Fits

- The technique we have used so far is called "Linear Least Squares".
- You might think that the word "Linear" refers to the fact that the equation of the straight line is linear in x.
- This is not the reason.
- The "Linear" refers to the fact that the equation is linear in the determinable parameters m and b.

$$\hat{y} = mx + b$$

Linear means raised to the power one.

Linear Least Squares Regressions/Fits

- The following equations are linear in the parameters to be determined (m's and b's):

$$\hat{y} = mx + b$$

$$\hat{y} = m_1 x + m_2 x^2 + m_3 x^3 + b$$

$$\hat{y} = m_e e^x + m_s \sin x + b$$

Can use linear least squares

- These following equations are examples where the parameters are <u>not</u> linearly related to the dependent variable:

Cannot use linear least squares

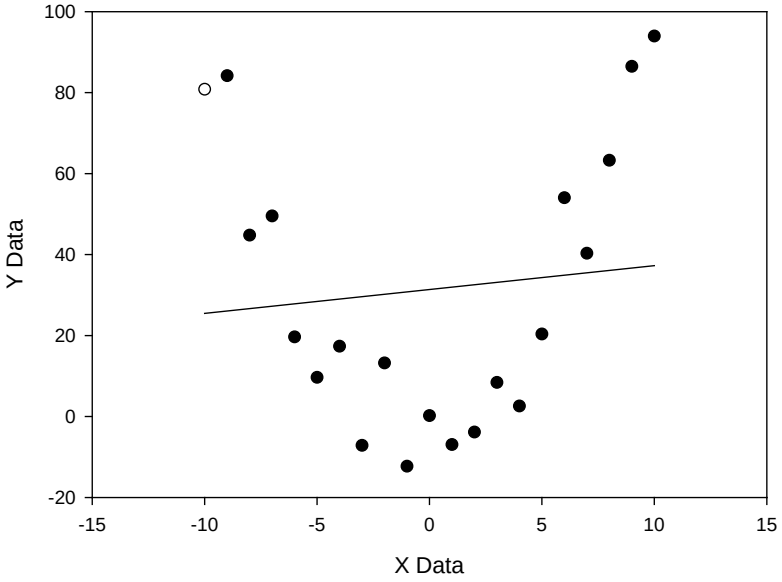
$$\hat{y} = m^x + \log b$$

$$\hat{y} = 3e^{-mx} + 15 \sin bx$$

Linear means raised to the power one.

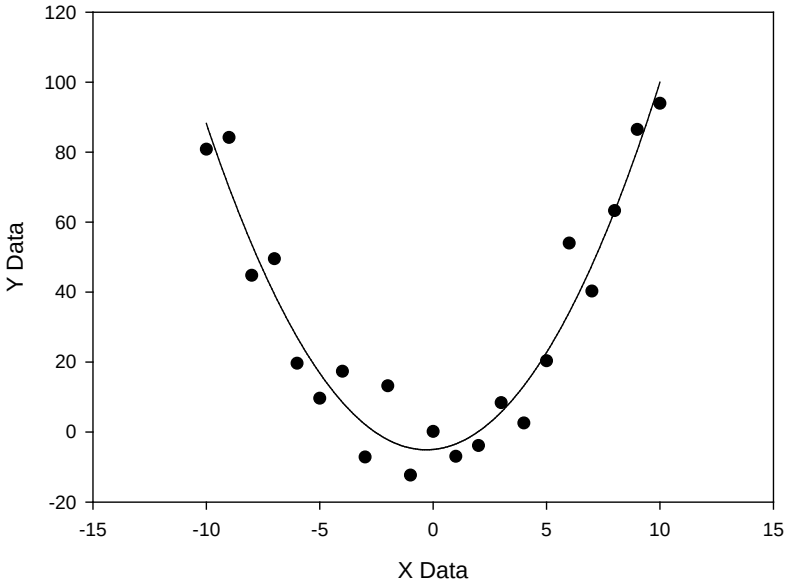
Correlation Coefficient and the Coefficient of Determination for Linear Least Squares

$$y = x^2 + \text{noise}$$



- In the top plot $r = 0.11$, which means $r^2 = 0.012$, which means 1% of the variability in the data can be attributed to a linear relationship between x and y.
- In the bottom plot the data are fitted to

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

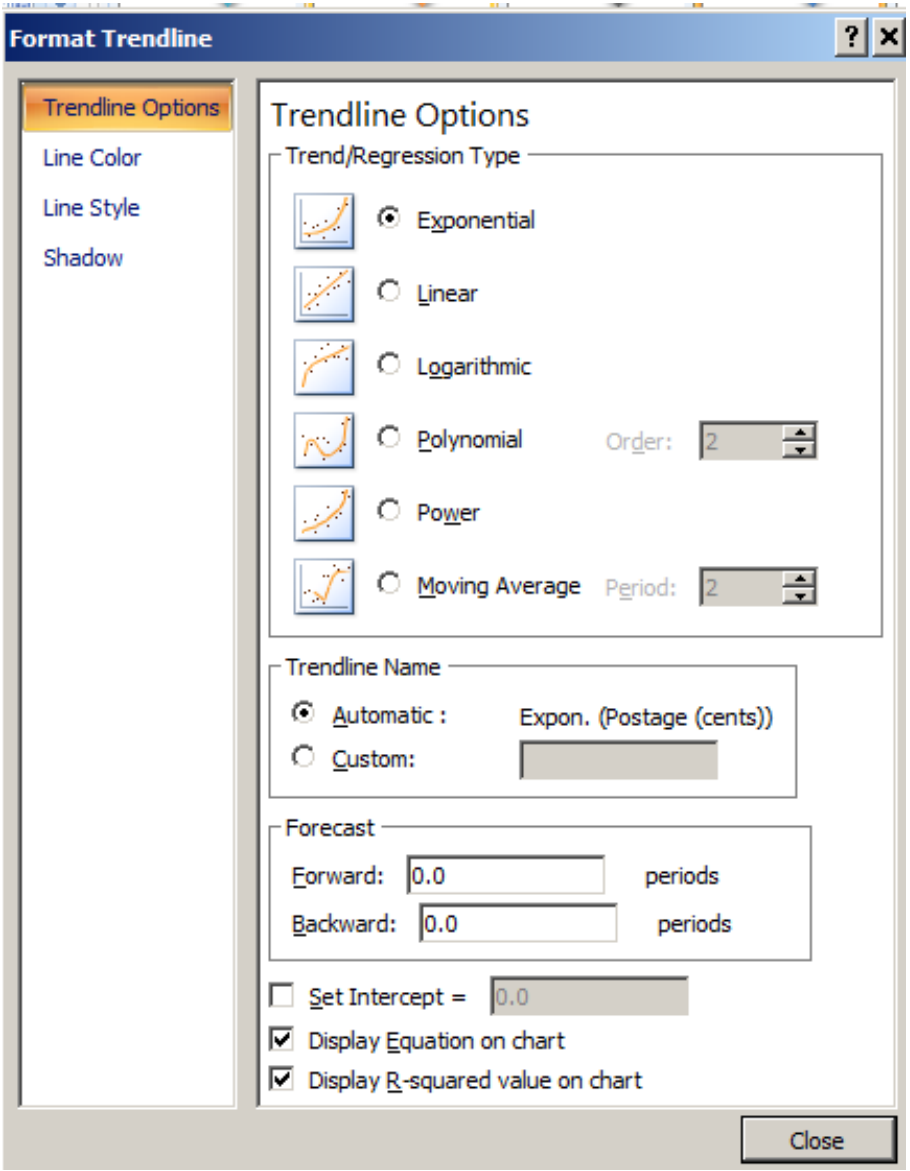


- The Coefficient of Determination for the fit was $r^2 = 0.929$
- Or, 92.9% of the variability in the y-data can be attributed to the fitted relationship between x and y.

Trend Lines

- Excel and most other plotting programs have a ‘trend line’ option.
- When the trend line is a non-linear function, like an exponential for example, the program `manipulates` the data so that linear least squares can be done (See page 612 in the 6th ed of the text book: Introduction to Engineering).
- This is okay for a trend line that is used only as a guide for the eye.
- But, it is absolutely not okay for serious data analysis.

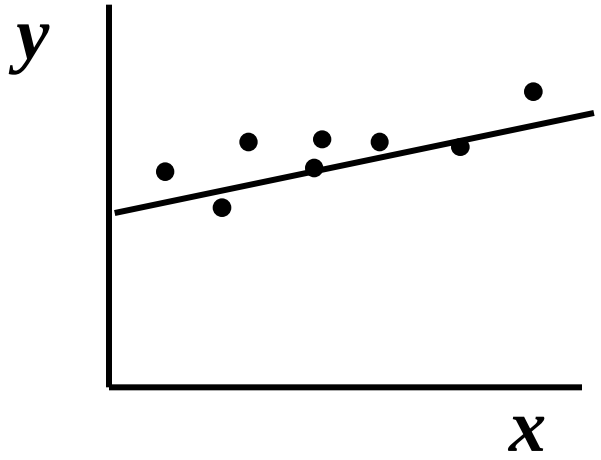
Trend Lines



Be careful with ‘trend lines’; they should not be used for anything other than producing a guide for the eye, unless you know that the underlying assumptions of least-squares regressions are satisfied by your data!!!!!

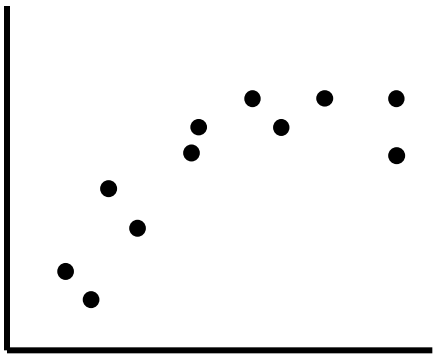
Assumptions

- Regression lines and tests of regression or correlation don't do you much good if the basic assumptions of linear regression are not met
- The first assumption for regression to a straight line is that the relation between *x* and *y* can be described by a straight line

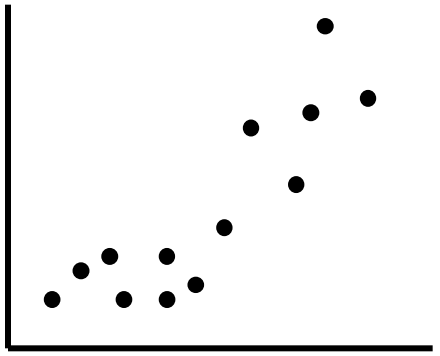


Assumptions

- If this assumption is not met the regression will still produce values for the slope and intercept, but these will not be interpretable
- Floor and ceiling effects often cause nonlinear relationships

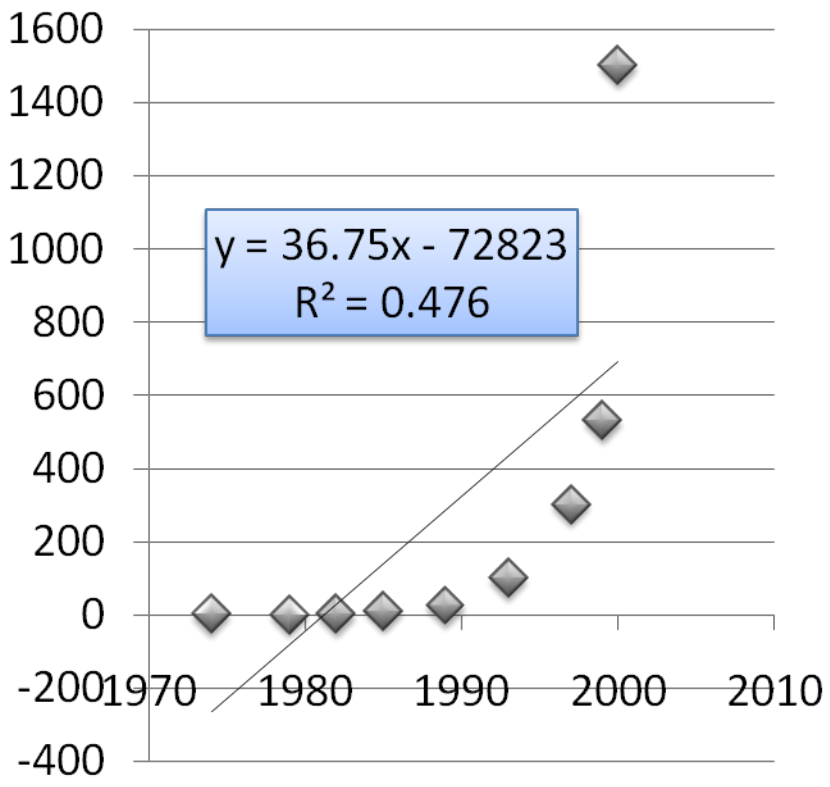


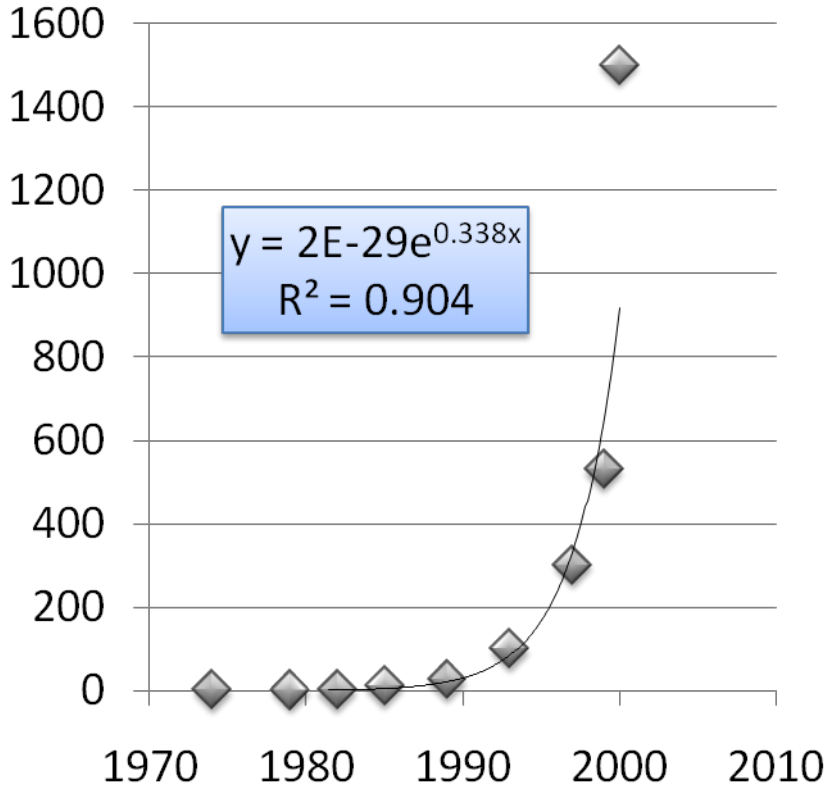
ceiling effect



floor effect

Linear and Exponential Trend Lines





Trend lines are useful as guides for the eye:
they provide a line showing the `stated` trend in the data,
but they cannot be relied upon for serious analysis.

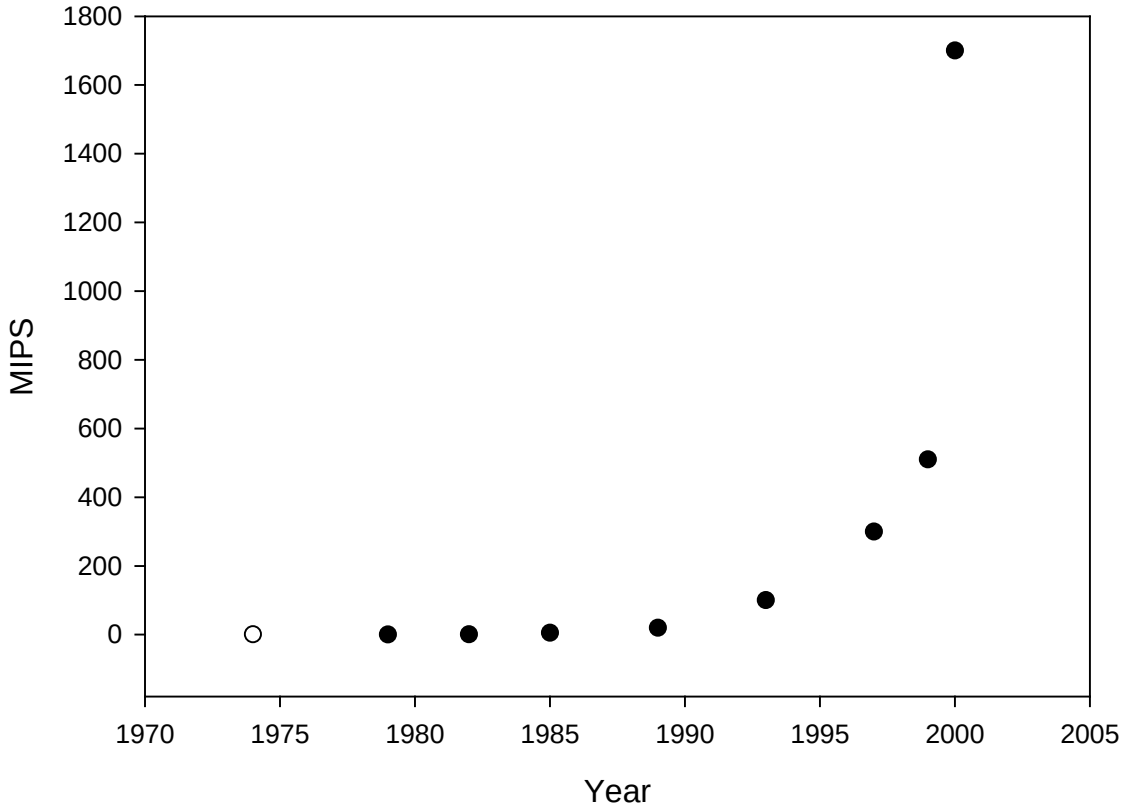
MIPS for Selected Intel Processors

Name	Year	MIPS
8080	1974	0.64
8088	1979	0.33
80286	1982	1
80386	1985	5
80486	1989	20
Pentium	1993	100
Pentium II	1997	300
Pentium III	1999	510
Pentium IV	2000	1700

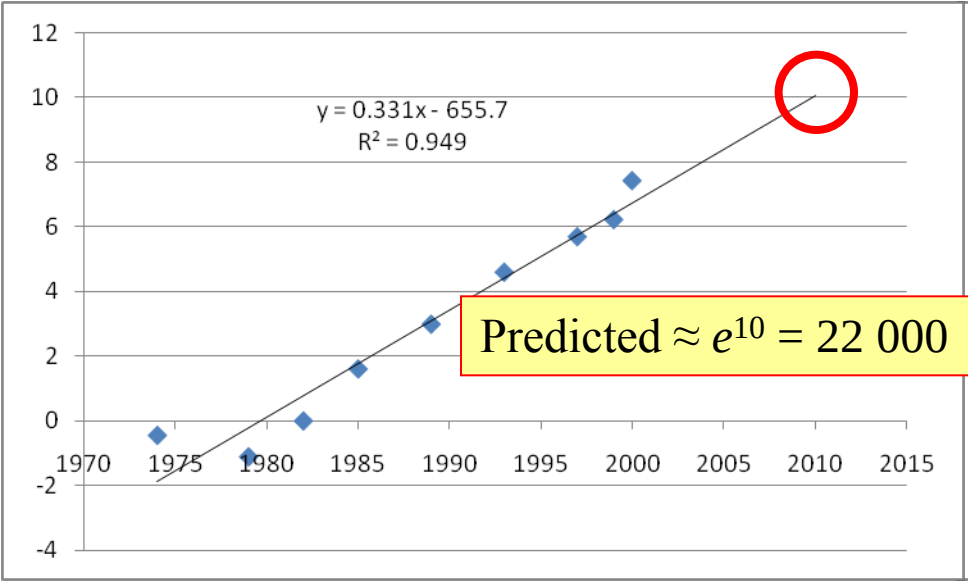
Task: Predict the speed (in MIPS) in 2010

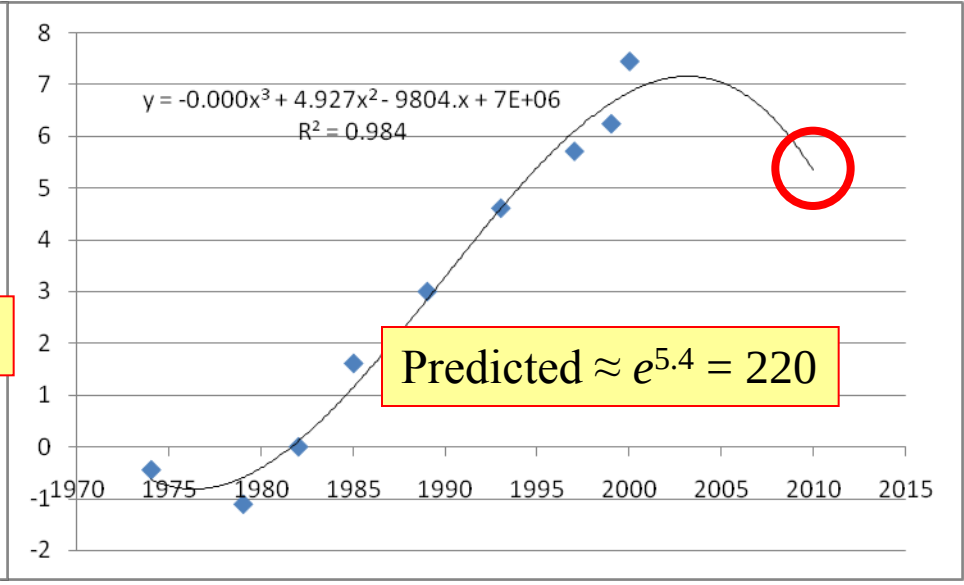
MIPS: Millions of Instructions Per Second

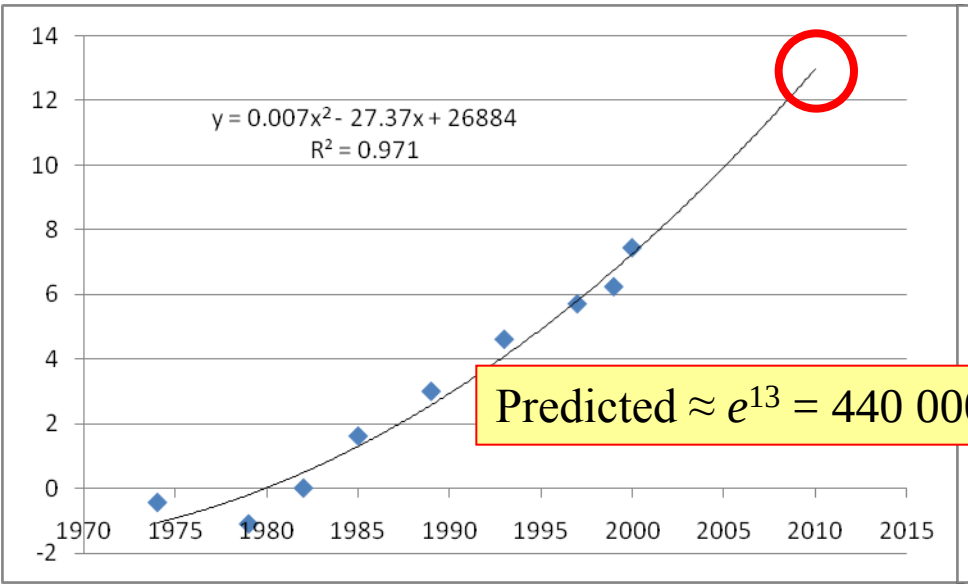
MIPS increases with time

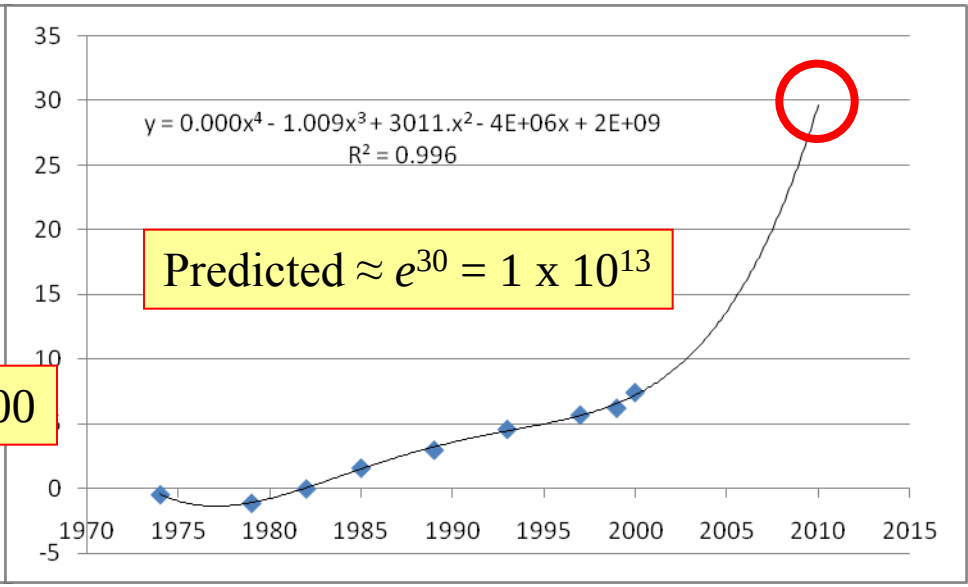


Moore's Law (Conjecture): computer speeds will increase exponentially with a doubling time of roughly 18 months.



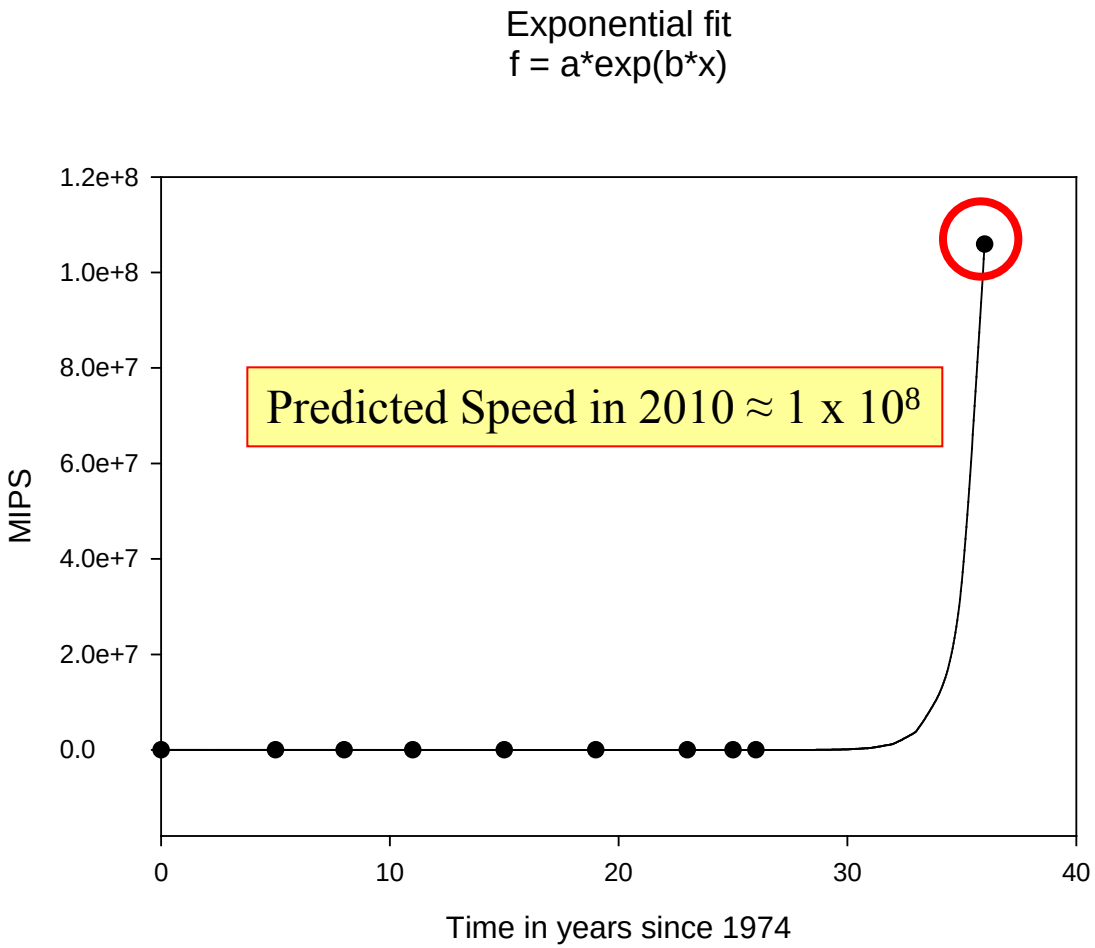






Vertical axis is natural log of MIPS

The same data, but fitted to an exponential directly gives:



Bee careful, this analysis is also not sufficient !!

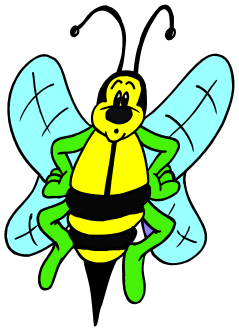
Never use Trend Lines for Serious Analysis !!!!

Be Careful

- For this reason, people are often careful when choosing their words, for instance,

  r^2 (explained variance) is "the amount of variance in *y* accounted for by a *linear relationship* between *x* and *y*"

- Often a nonlinear relationship will actually explain more of the overall variance than *you would estimate by simply squaring the linear correlation coefficient. Recall the parabola ($y=x^2$) fit a few slides back.*



Bee Careful

Competitive Interactions Between Neotropical Pollinators and Africanized Honey Bees

Abstract. *The Africanized honey bee, a hybrid of European and African honey bees, is thought to displace native pollinators. After experimental introduction of Africanized honey bee hives near flowers, stingless bees became less abundant or harvested less resource as visitation by Africanized honey bees increased. Shifts in resource use caused by colonizing Africanized honey bees may lead to population decline of Neotropical pollinators.*

SCIENCE, VOL 201, 15 SEPTEMBER 1978

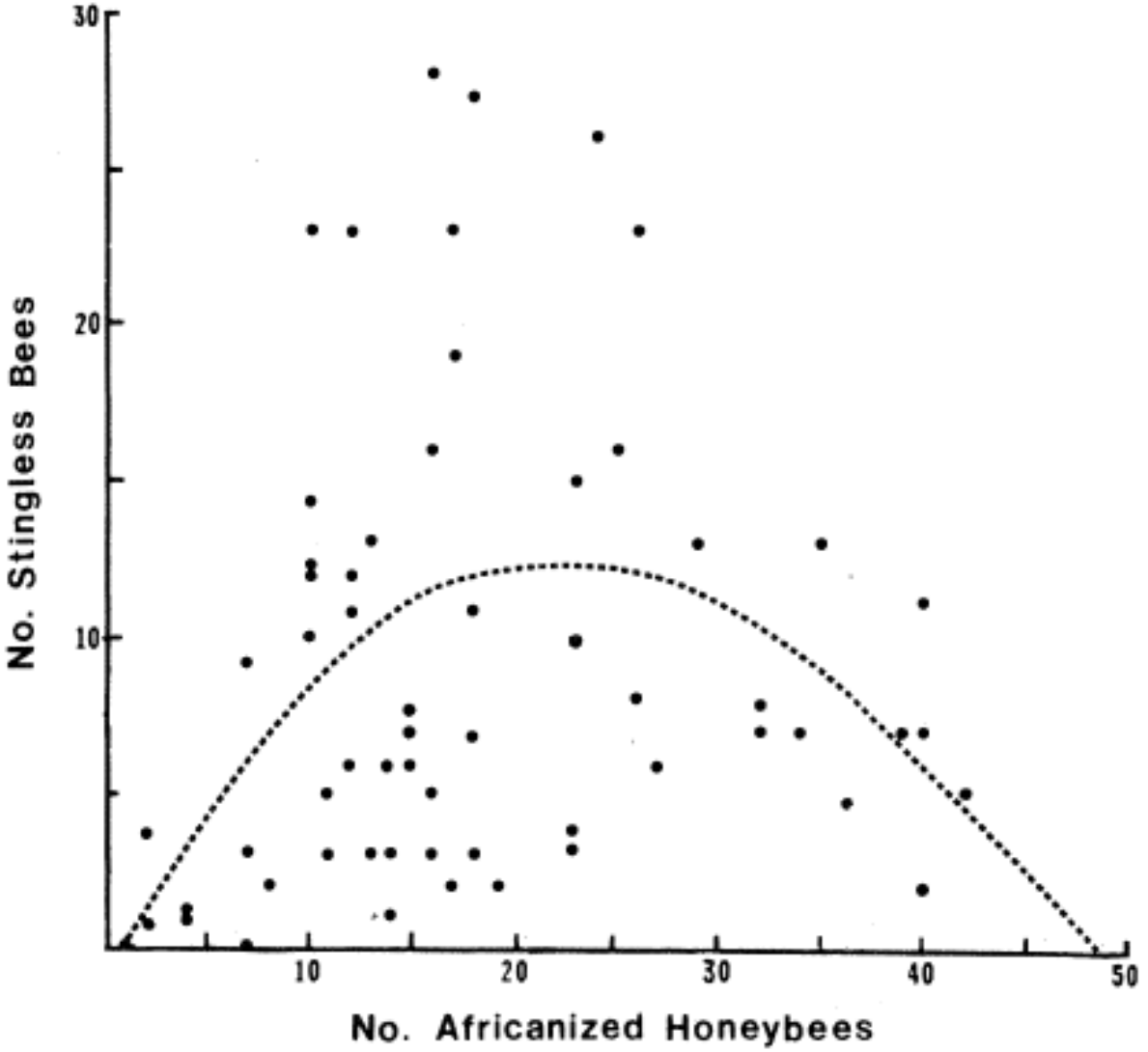


Fig. 1. The relations of Africanized and stingless (meliponine) bee abundances on flowering *Melochia villosa*. The dashed line is a quadratic polynomial (given by $y = -0.516 + 1.08x - 0.023x^2$) which gave the best fit to the points (*7*).

Curve-Fitting

The rather fanciful curve-fitting of Roubik (Reports, 15 Sept., p. 1030, Fig. 1) has prompted me to propose an alternative interpretation of his data (see below).

ROBERT M. HAZEN
Geophysical Laboratory,
Carnegie Institution of Washington,
Washington, D.C. 20018

- Bee careful using curve fitting programs:
  – it is easy to make mistakes!

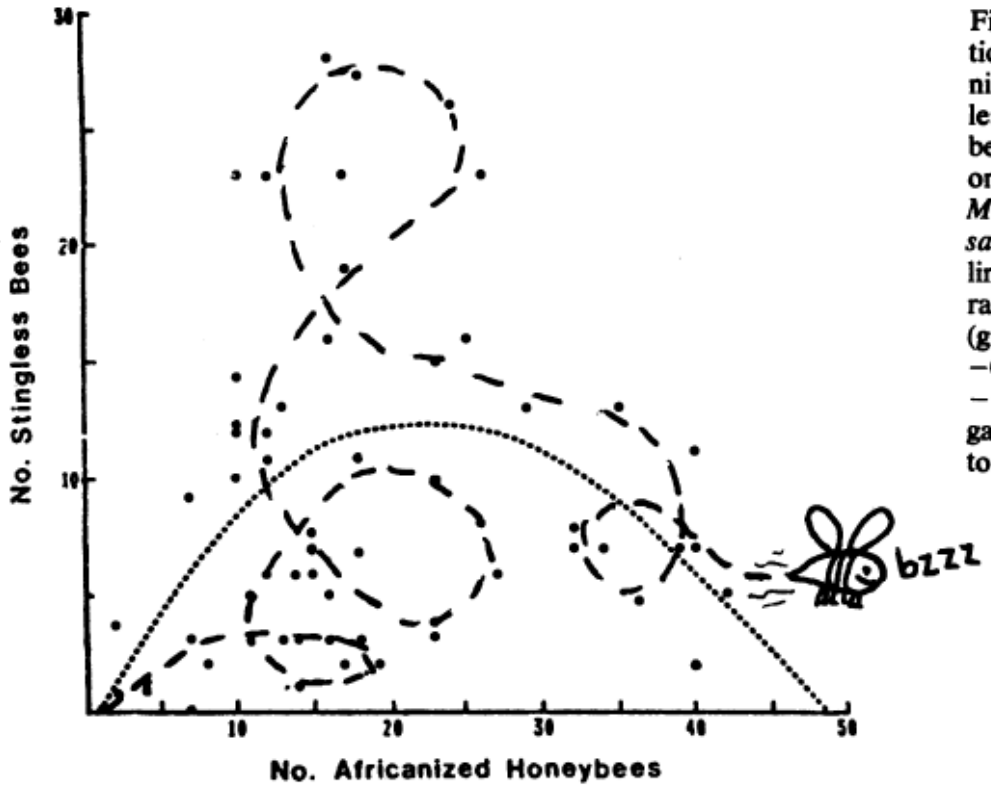


Fig. 1. The relations of Africanized and stingless (meliponine) bee abundances on flowering *Melochia villosa*. The dashed line is a quadratic polynomial (given by $y = -0.516 + 1.08x - 0.023x^2$) which gave the best fit to the points (7).