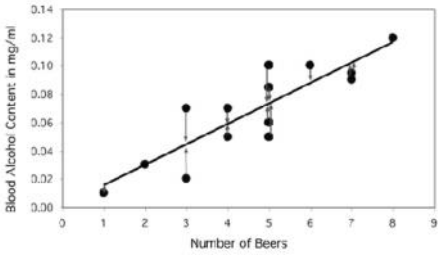
The Regression Line

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

$$b_1 = r\frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$



Example:

Here we have two quantitative variables for each of 16 students.

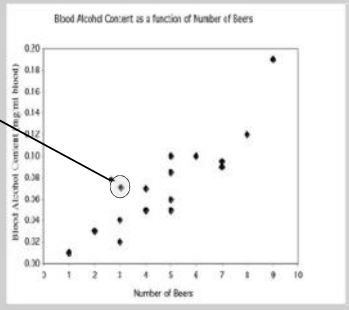
1. ***How many beers they drank, and***
2. ***Their blood alcohol level (BAC)***

We are interested in the relationship between the two variables: How is one affected by changes in the other one?

Student	Number of 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

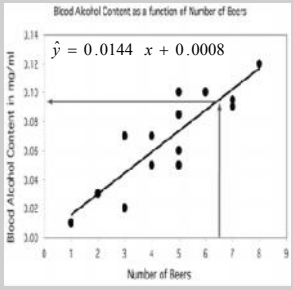
In a **scatterplot** one axis is used to represent each of the variables, and the data are plotted as points on the graph.

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



Making predictions: interpolation

The equation of the least-squares regression allows you to predict *y* for any *x* **within the range studied.** This is called **interpolating**.



Nobody in the study drank 6.5 beers, but by finding the value of $\hat{y}$ from the regression line for $x = 6.5$, we would expect a blood alcohol content of 0.094 mg/ml.

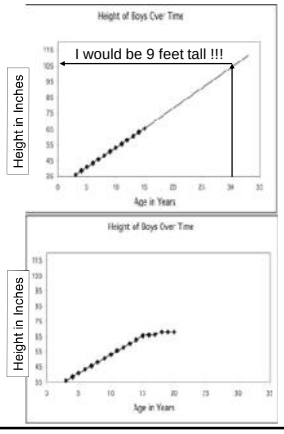
$$\hat{y} = 0.0144 * 6.5 + 0.0008$$

$$\hat{y} = 0.936 + 0.0008 = 0.0944 mg/ml$$

Making predictions: extrapolation

Extrapolation is the use of a regression line for predictions outside the range of *x* values used to obtain the line.

This can be a very silly thing to do, as seen here.

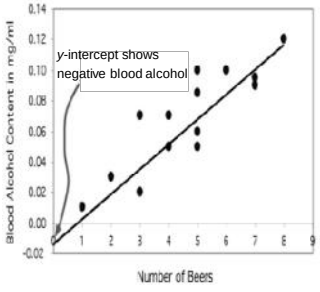


The *y*-intercept

Sometimes the *y*-intercept is **not possible**. Here we have negative blood alcohol content, which makes no sense…

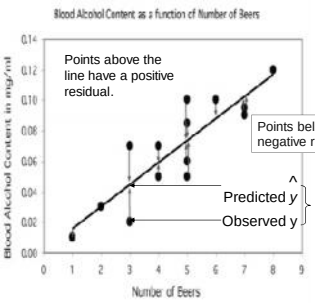
But the negative value is appropriate for the equation of the regression line.

There is a lot of scatter in the data and the line is just an estimate.



Residuals

The distances from each point to the least-squares regression line give us potentially useful information about the contribution of individual data points to the overall pattern of scatter.



These distances are called **"residuals."**

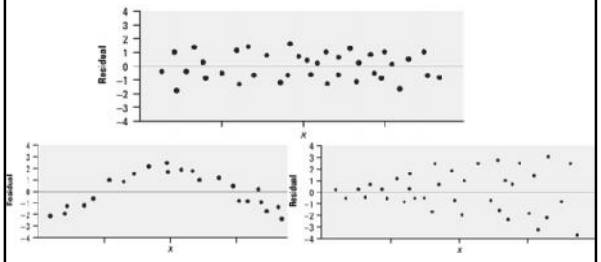
The sum of these residuals is always 0.

$$\text{dist.}\ (y-\hat{y}) = \text{residual}$$

Residual Plot

Residual Plots

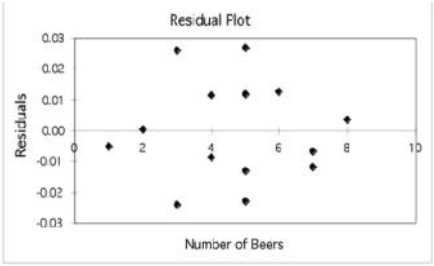
A **residual plot** is a scatterplot of the regression residuals against the explanatory variable (or equivalently, against the predicted y-values). Residual plots help us assess how well a regression line fits the data.

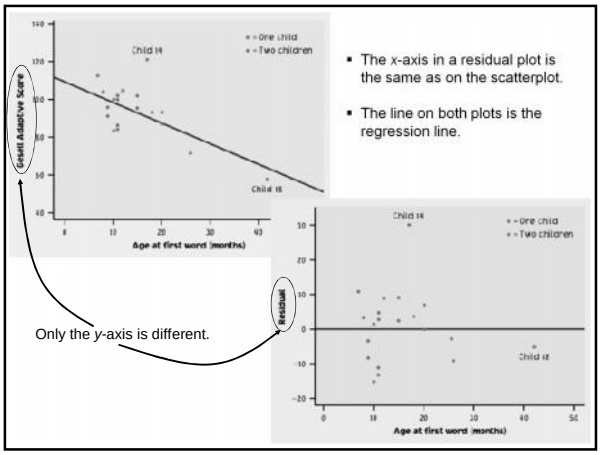


Residual plots

Residuals are the distances between *y*-observed and *y*-predicted. We plot them in a **residual plot**.

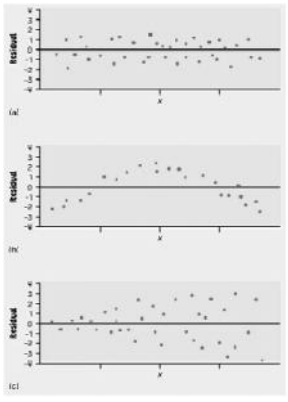
If residuals are scattered randomly around 0, chances are your data fit a linear model, were normally distributed, and you didn't have outliers.





- The *x*-axis in a residual plot is the same as on the scatterplot.
- The line on both plots is the regression line.

Only the *y*-axis is different.



Residuals are randomly scattered—good!

A curved pattern—means the relationship you are looking at is not linear.

A change in variability across plot is a warning sign. You need to find out why it is and remember that predictions made in areas of larger variability will not be as good.

Always plot your data!

The correlations all give $r \approx 0.816$, and the regression lines are all approximately $\hat{y} = 3 + 0.5x$. For all four sets, we would predict $\hat{y} = 8$ when $x = 10$.

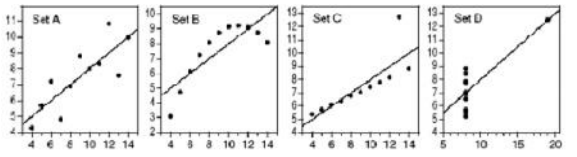
Table 2.5 Four data sets for exploring correlation and regression

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.

Always plot your data!

However, making the scatterplots shows us that the correlation/ regression analysis is not appropriate for all data sets.



Moderate linear association; regression OK.

Obvious nonlinear relationship; regression inappropriate.

One point deviates from the (highly linear) pattern of the other points; it requires examination before a regression can be done.

Just one very influential point and a series of other points all with the same *x* value; a redesign is due here…

The Coefficient of Determination: R^2

Coefficient of determination, R^2

- Gives the fraction of the variability of ***y*** accounted for the LSRL on ***x***
- Is an overall measure of how successful the regression is in linearly relating ***y*** to x.
- Say ***"____% of the variability of y (context) is accounted for by the variation in x (context)."***

Meaning of Coefficient of determination

- The determination can be thought of as a percent. Roughly speaking, it tells how many of the points of data fall within the results of the line formed by the **regression equation**. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted. If the coefficient is 0.80, then 80% of the points should fall within the regression line. Values of 1 or 0 would indicate the regression line represents all or none of the data, respectively. A higher coefficient is an indicator of a better goodness of fit for the observations.
- Its main purpose is the prediction of future outcomes on the basis of other related information.
- It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model.

Usefulness of Coefficient of determination

- The usefulness of R^2 in Statistics is its ability to determine the likelihood of future events falling within the predicted outcomes. The idea is that if more samples were added, the coefficient would show the probability of the new point falling on the line. Because it is possible to gain more samples, it is possible to test the viability of determination as a prediction tool.
- Similar to correlation, it should be noted that even if there is a strong connection between the two variables, ***determination does not prove causality***. For instance, a study on birthdays may show a large number of birthdays occur specifically within a time frame of one or two months. This does not mean that the passage of time or the change of seasons causes pregnancy.

Coefficient of determination, R^2

- The usefulness of R^2 in Statistics is its ability to determine the likelihood of future events falling within the predicted outcomes. The idea is that if more samples were added, the coefficient would show the probability of the new point falling on the line. Because it is possible to gain more samples, it is possible to test the viability of determination as a prediction tool.
- Similar to correlation, it should be noted that even if there is a strong connection between the two variables, ***determination does not prove causality***. For instance, a study on birthdays may show a large number of birthdays occur specifically within a time frame of one or two months. This does not mean that the passage of time or the change of seasons causes pregnancy.

Interpretation of R^2

- Consider Tampa sales example. From printout, $R^2 = 0.9453$.
- ***Interpretation:*** **94% of the variability observed in sale prices can be explained by assessed values of homes.**
- Thus, the assessed value of the home contributes a lot of information about the home's sale price.

Creating a Residual Plot

- Paste FEMUR and HUMER into L_1 & L_2
- Define L_3 with the formula $L_2 - Y_1(L_1)$
  - Notice that this is just y – y-hat, so L_3 now contains all the residuals of these data
  - This assumes Y_1 is the LSRL equation
- Set up Stat Plot 3 as a scatterplot of L_1 and L_3
- Turn off Plot 1 (the data) and Y_1 (the LSRL), turn on Plot 3 and tap Zoom-9 to see Plot 3
  - You are seeing a scatterplot of the residuals vs the explanatory variable
  - This is called a ***residual plot***

What Does it Mean?

- If the data were perfectly linear, the residuals would all be zero.
  - Then the residual plot would be points exactly on the x axis
- If the data miss the LSRL due to random variation, we expect the residuals to be randomly distributed.
  - Some will be positive, some negative
  - There should be no apparent pattern
- If the data miss the LSRL due to some curve in the data, the residuals will show a pattern.
  - Where the LSRL goes through the data, residuals will be small
  - Elsewhere the residuals will be large
  - The result is a residual plot that shows a curved pattern
- Conclusion: If the residual plot looks like a patternless "cloud of points" then the LSRL is a good model of the data

A Calculator Trick

- In your STAT : EDIT screen, scroll right so you can see L_3 and L_4
- Paste a list called RESID into L_4
  - It's in your calculator already!
- Note that RESID is identical to L_3
- Every time you run LinReg, the calculator computes all the residuals and places them in RESID
- Edit Stat Plot 3 so that the Ylist is based on RESID instead of L_3
  - Now leave it that way!
  - After every LinReg, Plot 3 will always show you the residual plot.