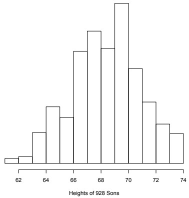
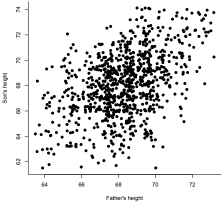
# Prediction is a key task of statistics

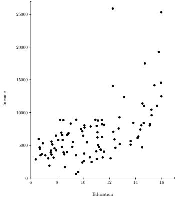


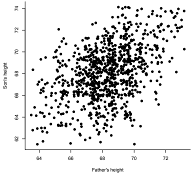
Predict the height of a son who is chosen at random from 928 sons. The average height of sons, 68.1 in, is the 'best' predictor.



Predict the height of a son whose father is 72 in tall. This additional information about the father should allow us to make a better prediction. Regression does just that.

# The correlation coefficient





The scatterplot visualizes the relationship between two quantitative variables. It may have a direction (sloping up or down), form (a scatter that clusters around a line is called *linear*) and strength (how closely do the points follow the form?).

If the form is linear, then a good measure of strength is the **correlation coefficient r**: Our data are $(x_i, y_i),\ i = 1, \ldots, n$.

$$r \;=\; \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

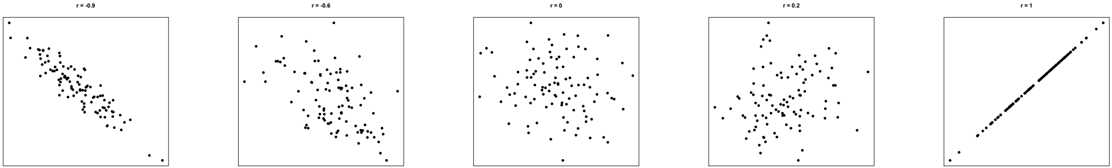
(divide by $n-1$ instead of $n$ if this is also done for the standard deviations $s_x, s_y$).

# Correlation measures linear association

A numerical summary of these pairs of data is given by: $\bar{x}, s_x, \bar{y}, s_y, r$.

As a convention the variable on the horizontal axis is called **explanatory variable** or **predictor**, the one on the vertical axis is called **response variable**.

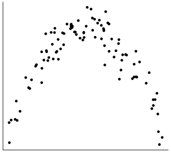
$r$ is always between $-1$ and $1$. The sign of $r$ gives the direction of the association and its absolute value gives the strength:



Since both $x$ and $y$ were standardized when computing $r$, $r$ has no units and is not affected by changing the center or the scale of either variable.

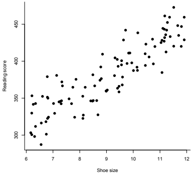
# Correlation measures linear association

Keep in mind that $r$ is only useful for measuring *linear* association:



$$r = 0$$

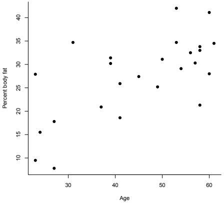
Also remember that correlation does not mean causation:

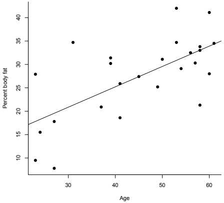


Among school children there is a high correlation between shoe size and reading ability. Both are driven by the *lurking variable* 'age'.

# The regression line

If the scatterplot shows a linear association, then this relationship can be summarized by a line.





To find this line for $n$ pairs of data $(x_1, y_1), \ldots, (x_n, y_n)$, recall that the equation of a line produces the y-value $\hat{y}_i = a + bx_i$. The idea is to choose the line that minimizes the sum of the squared distances between the observed $y_i$ and the $\hat{y}_i$. In other words, find $a$ and $b$ that minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

# The method of least squares

For $n$ pairs of data $(x_1, y_1), \ldots, (x_n, y_n)$, find $a$ and $b$ that minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \; = \; \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

This is the **method of least squares**. It turns out that $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$. This line $\hat{y} = a + bx$ is called the **regression line**.

There is another interpretation of the regression line:
it computes the average value of $y$ when the first coordinate is near $x$.

Remember that often times an average is the 'best' predictor. This shows how the regression line incorporates the information given by $x$ to produce a good predictor of $y$.

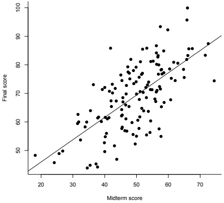
# Regression to the mean

The main use of regression is to predict $y$ from $x$:
Given $x$, predict $y$ to be $\hat{y} = a + bx$.

The prediction for $y$ at $x = \bar{x}$ is simply $\hat{y} = \bar{y}$.

But $b = r\frac{s_y}{s_x}$ means that if $x$ is one standard deviation $s_x$ above $\bar{x}$, then the predicted $\hat{y}$ is only $r\, s_y$ above $\bar{y}$.

Since $r$ is between $-1$ and $1$, the prediction is 'towards the mean': $\hat{y}$ is fewer standard deviations away from $\bar{y}$ than $x$ is from $\bar{x}$.



# Regression to the mean

This is called **regression to the mean** (or: the **regression effect**). In can be observed in data whose scatter is football-shaped such as the exam scores: In such a test-retest situation, the top group on the test will drop down somewhat on the retest, while the bottom group moves up.

A heuristic explanation is this: To score among the very top on the midterm requires excellent preparation as well as some luck. This luck may not be there any more on the final exam, and so we expect this group to fall back a bit.

This effect is simply a consequence of there being a scatter around the line. Erroneously assuming that this occurs due to some action (e.g. 'the top scorers on the midterm slackened off') is the **regression fallacy**.

# Predicting $y$ from $x$ and $x$ from $y$

If we are given $x$, then we use the regression line $\hat{y} = a + bx$ to predict $y$.
To find this regression line we need only $\bar{x}, \bar{y}, s_x, s_y$ and $r$.

We can use software to compute this line, e.g. 'lm' in **R**, but it can also be done quickly by hand:

$\overline{\text{midterm}} = 49.5, \overline{\text{final}} = 69.1, s_{mid} = 10.2, s_{final} = 11.8, r = 0.67.$
Predict the final exam score of a student who scored 41 on the midterm.

# Predict $x$ from $y$

Predict the midterm score of a student who scored 89 on the final.

When predicting $x$ from $y$ **it is a mistake** to use the regression line $\hat{y} = a + bx$, derived for regressing $y$ on $x$, and solve for $x$. This is because regressing $x$ on $y$ will result in a **different regression line**.

To avoid confusing these, always put the predictor on the x-axis and proceed as on the previous slide.

# Normal approximation in regression

Regression requires that the scatter is football-shaped. Then one may use normal approximation for the $y$-values conditional on $x$. That is, the observations whose first coordinate is near that $x$ have $y$-values that approximately follow the normal curve.

To standardize, subtract off the predicted value $\hat{y}$, then divide by $\sqrt{1 - r^2} \times s_y$.

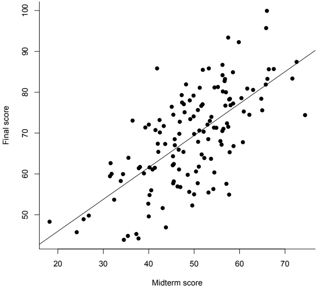
Among the students who scored around 41 on the midterm, what percentage scored above 60 on the final?

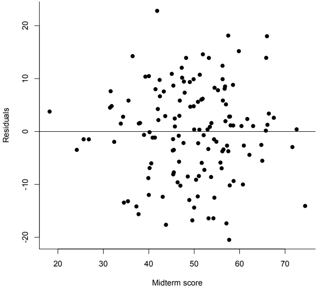
# Residuals

The differences between observed and predicted y-values are called **residuals**:

$$e_i \;=\; y_i - \hat{y}_i, \qquad i = 1, \ldots, n$$

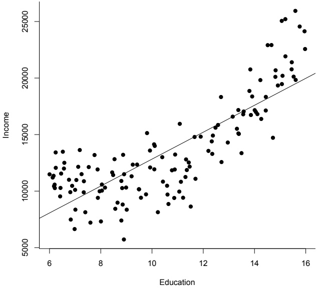
Residuals are used to check whether the use of regression is appropriate. The **residual plot** is a scatterplot of the residuals against the x-values. It should show an unstructured horizontal band.

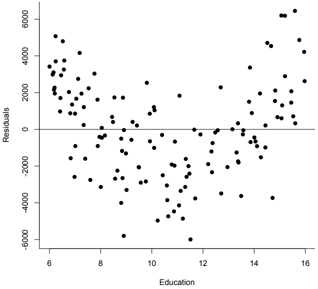




# Residual plots

A curved pattern suggests that the scatter is not linear:

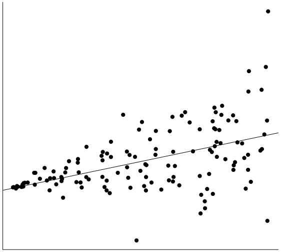


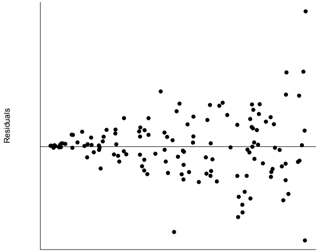


But it may still be possible to analyze these data with regression! Regression may applicable after **transforming** the data, e.g. regress $\sqrt{\text{income}}$ or log(income) on Education.

# Transformations of the variables

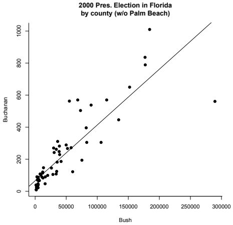
Another violation of the football-shaped assumption about the scatter arises if the scatter is **heteroscedastic**:

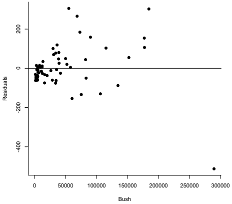




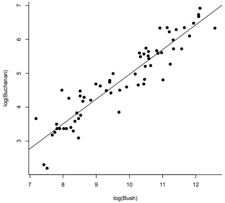
A transformation of the y-variables may produce a **homoscedastic** scatter, i.e. result in equal spread of the residuals across $x$. (However, it may also result in a non-linear scatter, which may require a second transformation of the x-values to fix!)

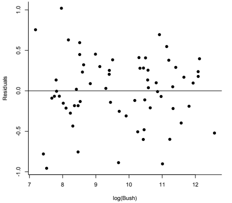
# Transformation of the variables





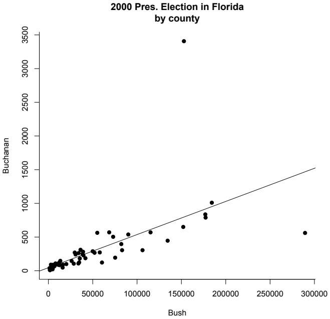
The residual plot looks heteroscedastic. Taking log of both variables produces a residual plot that is very satisfactory:

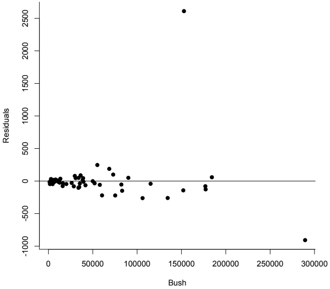




# Outliers

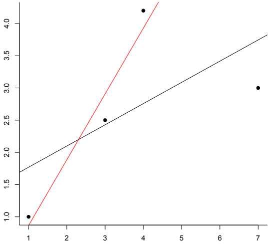
Points with very large residuals (**outliers**) should be examined: they may represent typos or interesting phenomena.





# Leverage and influential points

A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



Whether it does change the line a lot (→ **influential point**) or not can only be determined by refitting the regression without the point. An influential point may have a small residual (because it is influential!), so a residual plot is not helpful for this analysis.

# Some other issues

- Avoid predicting $y$ by **extrapolation**, i.e. at x-values that are outside the range of the x-values that were used for the regression: The linear relationship often breaks down outside a certain range.
- Beware of data that are summaries (e.g. averages of some data). Those are less variable than individual observations and correlations between averages tend to overstate the strength of the relationship.
- Regression analyses often report 'R-squared': $R^2 = r^2$. It gives the fraction of the variation in the y-values that is explained by the regression line. (So $1 - r^2$ is the fraction of the variation in the y-values that is left in the residuals.)