1 Regression

Regression

Linear models

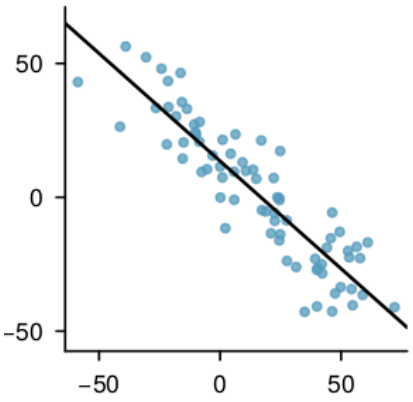
Problem: We have a set of points $\{(x_i, y_i)\}$. Given a new x value, we'd like to predict $\hat{y}$.

Linear model: We'll assume there exists a linear relationship $y = \beta_0 + \beta_1 x$ that offers a good approximation to the data.

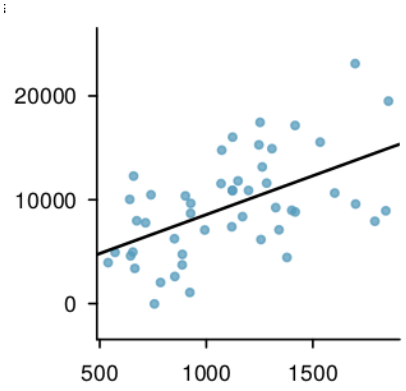
We call x the **explanatory** or **predictor** variable.

We call y the **response** variable.

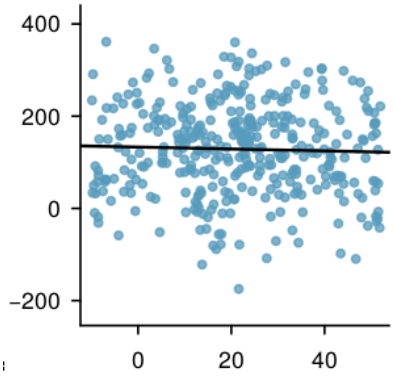
Example:



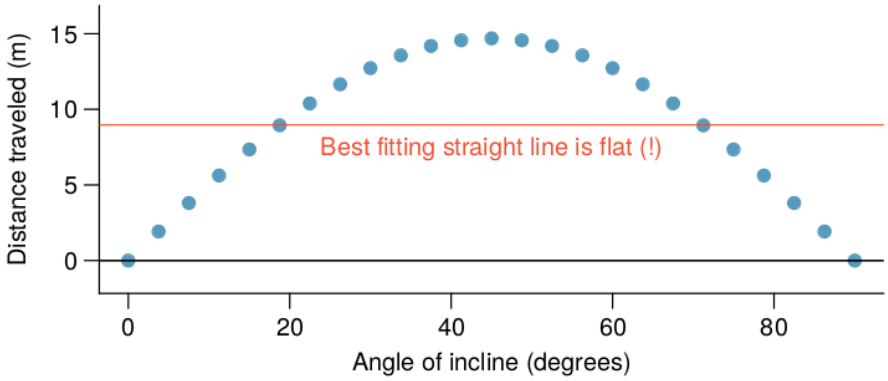
Example:



Example:



Example:



In the real world, there's nearly always noise.

Sometimes there are other lesser effects as well.

Talk about meaning of slope

Dangers of extrapolation. Example: global warming (a few data points in a few places at a few times)

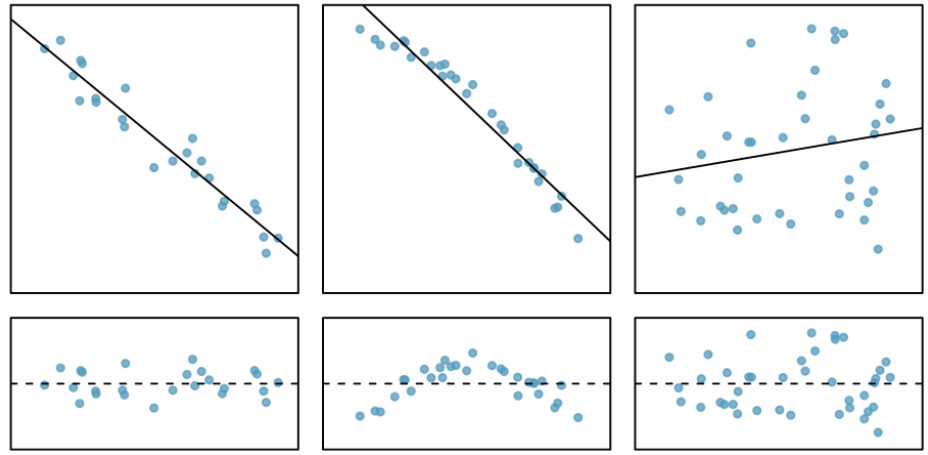
Residuals

What's left over.

$$\text{data} = \text{fit} + \text{residual}$$

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

Goal: small residuals!



Goal: small residuals.

$$\sum | e_i |$$

Goal: small residuals.

$$\sum e_i^2$$

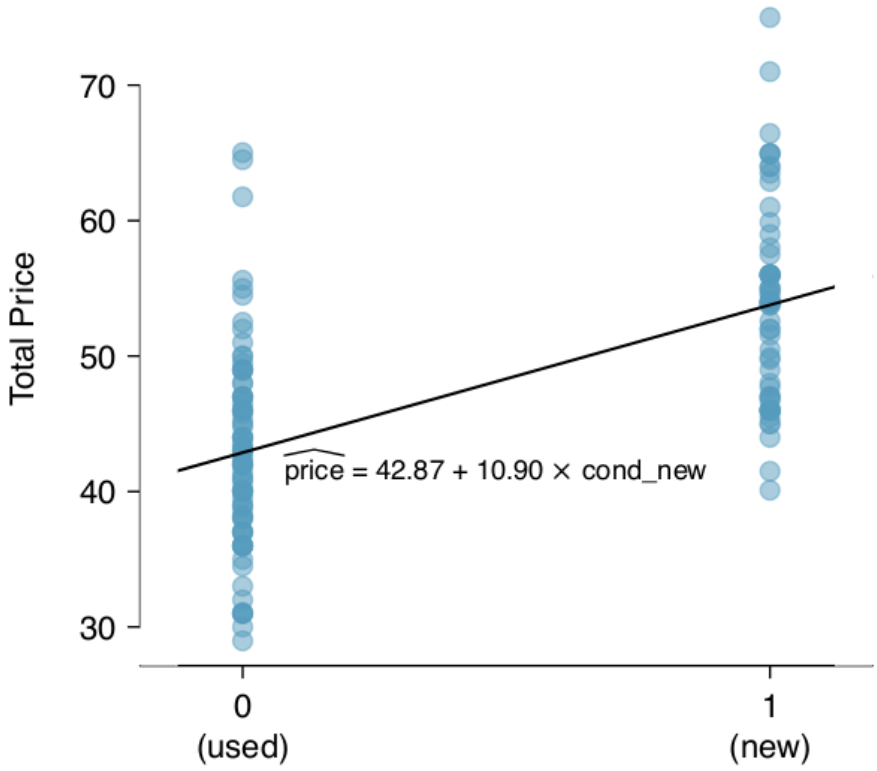
Residuals are what's left over after accounting for model fit.

A normal distribution of residuals is a good sign. And conversely.

Not rules: rule of thumb.

Time series often have important underlying structure. Correlation often doesn't model them well.

Categorical regression



Prices at ebay auctions.

TODO: Is this the same as line connecting means? If no, demonstrate.

Outliers

High leverage points fall far from the regression line.

Influential points make their leverage known.

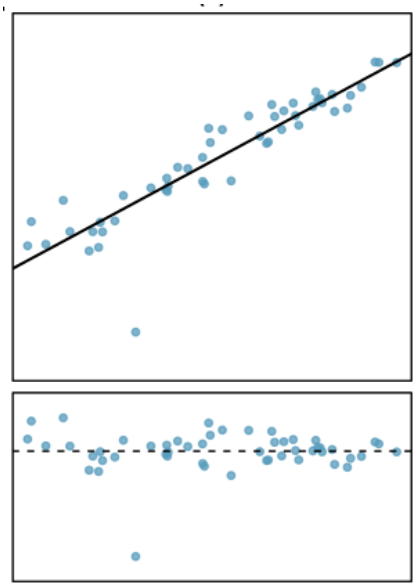
Points that fall farther from the regression line have more effect. We call them *high leverage* points.

If the effect is noticeable on the regression, we call it an *influential point*.

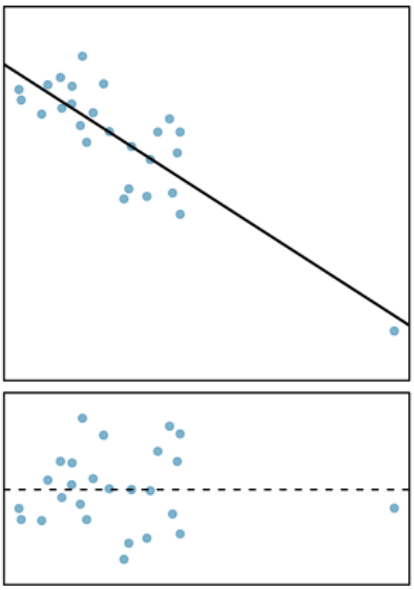
If a point, omitted, would fall much further from the regression line, it is certainly influential.

If not enough data points, they might be all or mostly influential!

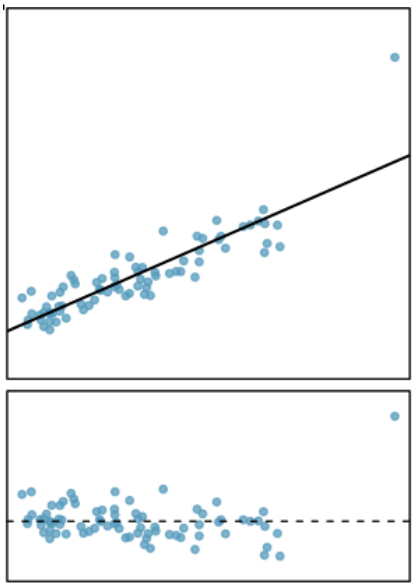
Outliers



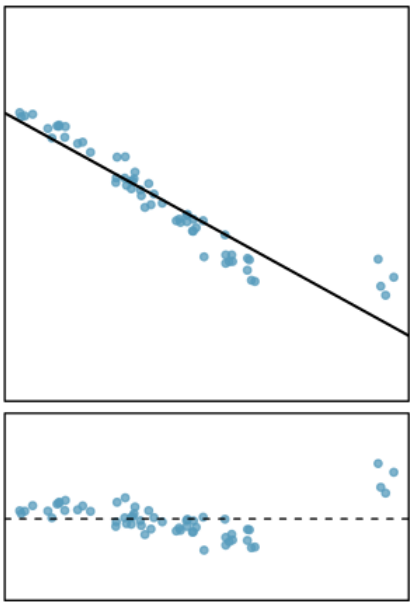
There is one outlier far from the other points, though it only appears to slightly influence the line.



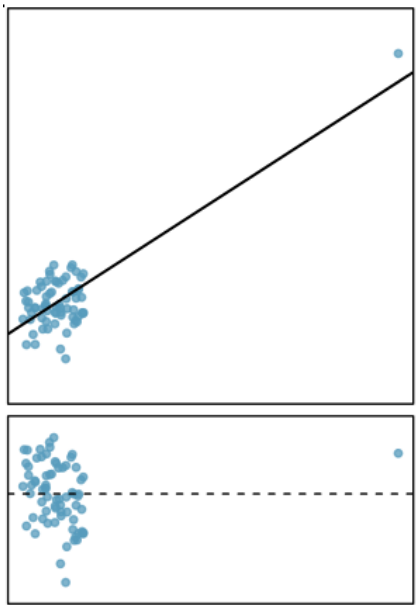
There is one outlier on the right, though it is quite close to the least squares line, which suggests it wasn't very influential.



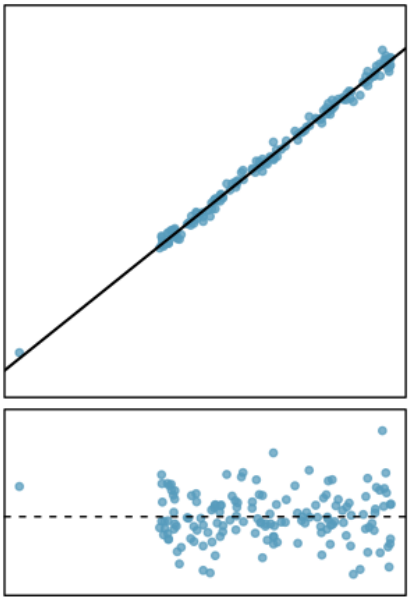
There is one point far away from the cloud, and this outlier appears to pull the least squares line up on the right; examine how the line around the primary cloud doesn't appear to fit very well.



There is a primary cloud and then a small secondary cloud of four outliers. The secondary cloud appears to be influencing the line somewhat strongly, making the least square line fit poorly almost everywhere. There might be an interesting explanation for the dual clouds, which is something that could be investigated.

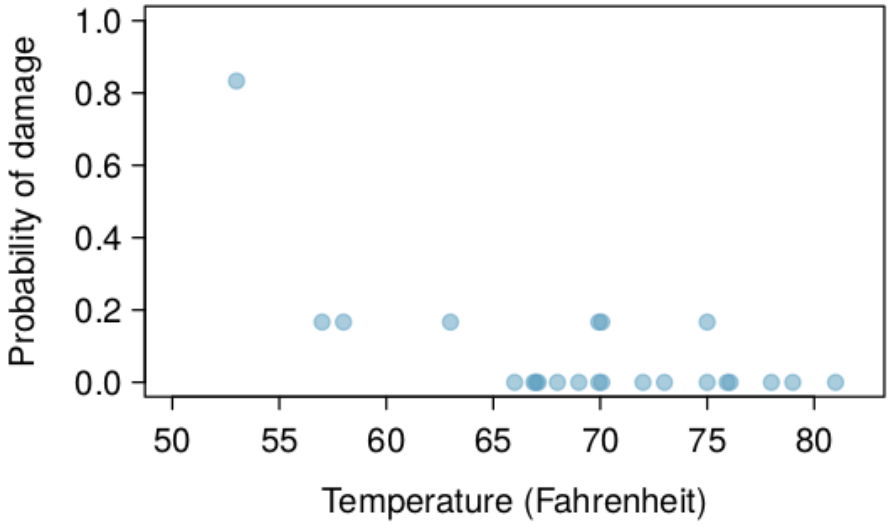


There is no obvious trend in the main cloud of points and the outlier on the right appears to largely control the slope of the least squares line.



There is one outlier far from the cloud, however, it falls quite close to the least squares line and does not appear to be very influential.

Don't ignore outliers.



Data about shuttle O-ring damage.

Correlation

Population correlation.

$$\rho_{X,Y} = \mathbf{Corr}(X,Y) = \frac{\mathbf{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbf{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Correlation measures dependence between two random variables.

Pearson's product-moment coefficient. Pearson's coefficient Pearson's correlation Correlation

(But there are others.) Pearson's correlation is sensitive to linear relationships.

If $\rho = 1$, we say perfect (increasing) correlation. If $\rho = -1$, we say perfect (decreasing) correlation or perfect anticorrelation.

Otherwise, indicates the degree of linear dependence.

Independent $\Rightarrow$ Pearson's correlation coefficient is zero. Converse not true. E.g., $y = x^2$ has zero correlation.

If X and Y are *jointly normal*, then zero correlation $\Rightarrow$ independent.

(*multivariate normal distribution*, *multivariate Gaussian distribution*)

r^2 describes amount of variation in the response explained by the least squares line. It's called the *coefficient of determination.* Sample correlation.

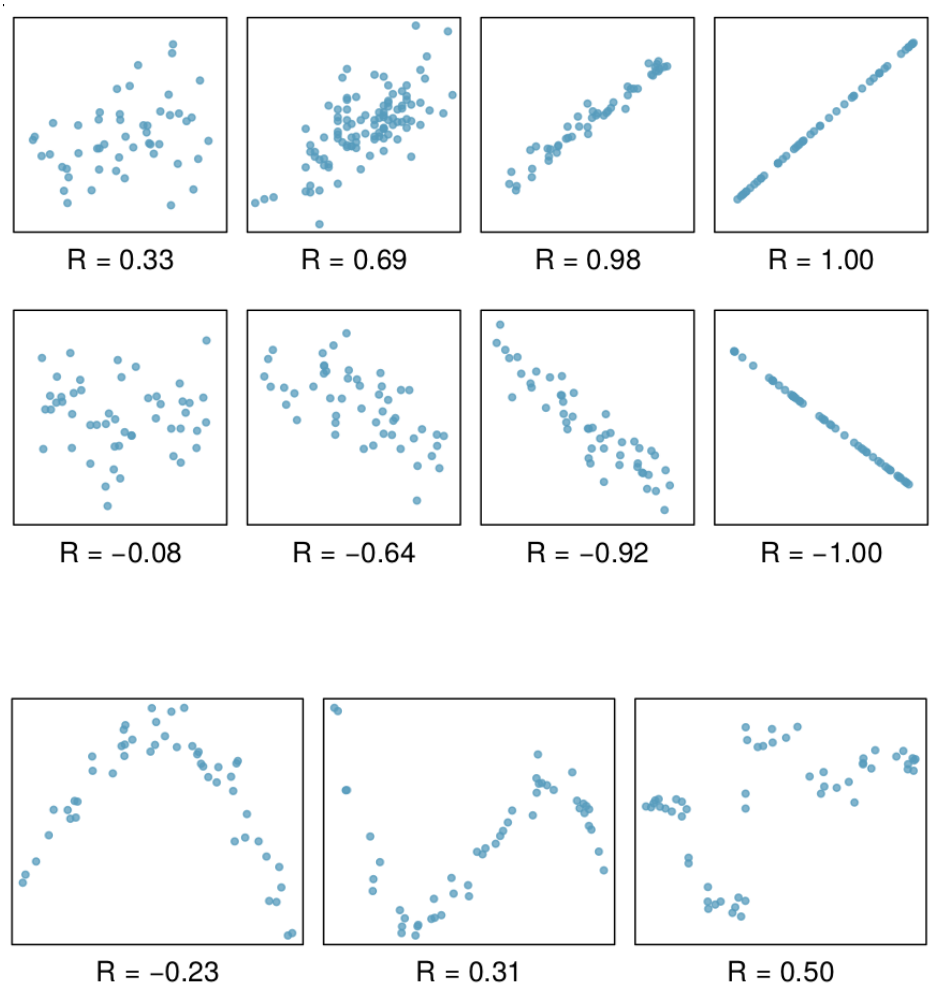
$$r_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right) \left(\frac{y_i - \overline{y}}{s_y} \right)$$

Coefficient of determination

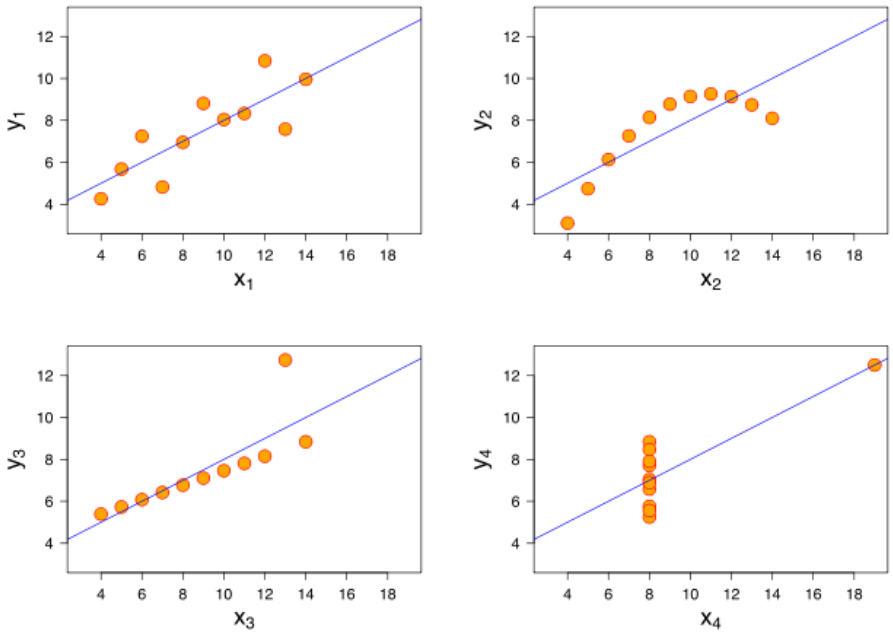
$$r^2 = \frac{\mathbf{Var}e_i}{\mathbf{Var}y_i}$$

This is (usually) true in one dimension under common conditions. In higher dimensions, we still call it r^2 but it's not longer the Pearson's correlation coefficient.

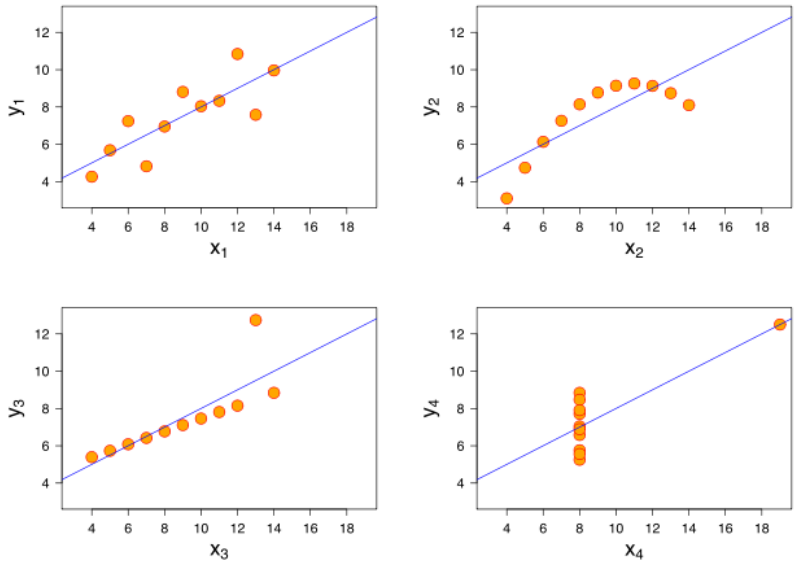
r^2 measures goodness of fit.



Anscombe's Quartet



Anscombe's Quartet



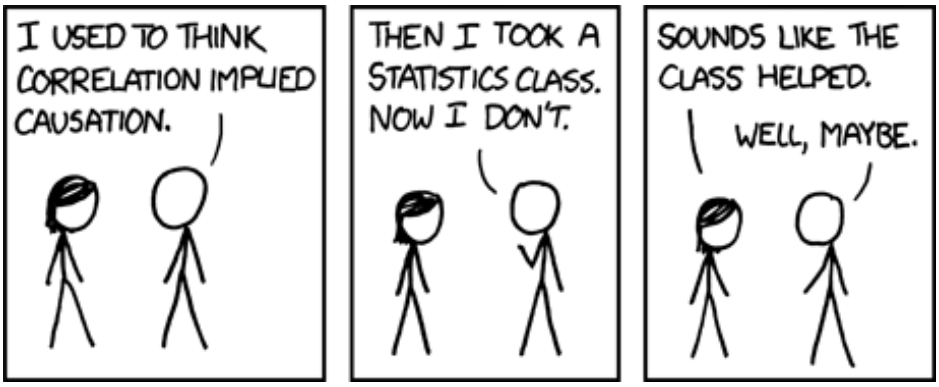
http://en.wikipedia.org/wiki/File:Anscombe\%27s_quartet_3.svg

Created by Francis Anscombe.

- mean = 7.5
- variance = 4.12
- correlation = 0.816
- regression line = $y = 3 + 0.5x$

Summary statistics don't replace visualising data.

Correlation does not imply causation



https://xkcd.com/552/

Tufte:

> "Empirically observed covariation is a necessary but not sufficient condition for causality."

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

Tufte:

> "Correlation is not causation but it sure is a hint."

http://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

In the xkcd, the salient point is that the character who didn't take the statistics class can't validly make that conclusion.

The character who took the statistics class, of course, probably can.

Code lab: linear regression

For each data set:

Tasks:

- Visualise the data
- Compute mean, variance, standard deviation, correlation
- Compute linear regression and visualise
- Compute residuals and visualise
- Discuss what the data tells you

Game: in pairs, set x to be a distribution for your neighbor.

Linear regression and inference

Remember about inference?

- Standard error
- p values

TODO. Talk about standard error and p values. Probably talk about t tests. Probably use sklearn and demo.

Multiple regression

One response $\hat{y}$ but many predictors $x_1, \ldots, x_n$.

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

We fit all predictors simultaneously. Fitting individually is not valid.

Logistic regression

Also: *logit regression*, *logit model*

Dependent variable (y) is binary.

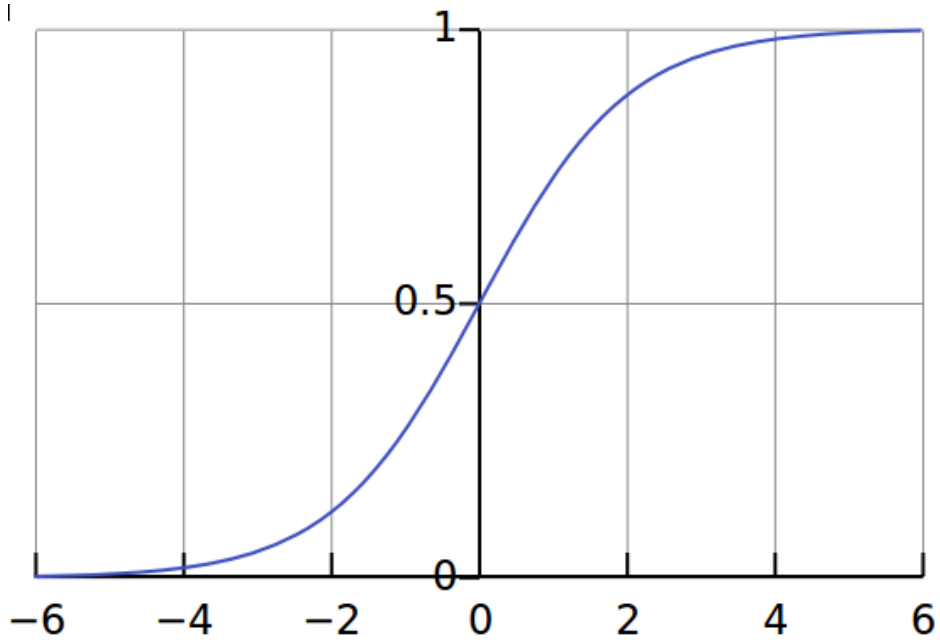
Dependent variable (y) is binary $\Rightarrow$ logistic regression is a classifier.

Logistic regression is an example of a *generalized linear model* (GLM).

Example: probability of dying given some non-boolean measure (e.g., size of tumour, time unconscious, etc.).

If y is binary, we can say *binary* or *binomial logistic regression*. If y may take on 3 or more discrete values, we speak of *multinomial logistic regression*.

Logistic regression used in activation functions in ANN's to keep responses bounded.



We compute residuals the same way: predicted value less observed value.

Note that $f(-x) = 1 - f(x)$.

And it satisfies this differential equation:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x) = f(x) \cdot (1 - f(x))$$

Integrating,

$$f(x) = \frac{e^x}{e^x + 1} = \frac{1}{1 + e^{-x}}$$

Fitting polynomials

- Under/over-fitting
- Test is not validation

There exists a unique polynomial of minimum degree that fits a given (finite) set of points.

$$S = \{(x_0, y_0), \ldots, (x_k, y_k)\}$$

Uniqueness proof is by counterexample: suppose $p(x)$ and $q(x)$ are minimal degree and fit S. Suppose further highest order coefficient is 1. Subtract. Zero at the $k+1$ points of S, so has degree at least $k+1$. But has degree at most k, since highest order term is now zero. So $p = q$.

The **Newton form** of the interpolation polynomial is a linear combination of Newton basis polynomials:

$$N(x) = \sum_{j=0}^{k} a_j n_j(x)$$

$$n_j(x) = \prod_i = 0^{j-1}(x - x_i)$$

for $j > 0$ and $n_0(x) \equiv 1$ and a_j is defined in terms of divided differences..

The **Lagrange form** is a linear combination of Lagrange basis polynomials:

$$L(x) = \sum_{j=0}^{k} y_j l_j(x)$$

$$l_j(x) = \prod_{m \in [0,k], m \neq j} \frac{x - x_m}{x_j - x_0}$$

Subject to Runge's phenomenon (oscillation near the interval edges). Higher degrees doesn't always improve accuracy.

We need a test set and a different validation set. (Typically, we leave some data out for validation.)

Barrier functions (e.g., sum of absolute values of coefficients) can help avoid over-fitting. There are other techniques. It depends on context.

Questions?

purple.com/talk-feedback

2 Break

Break