# DATA70121: Statistics and Machine Learning 1

## Lecture 5: Regression I

Diego Perez Ruiz
Department of Social Statistics
School of Social Sciences
University of Manchester

26 October 2023

# this lecture

**modelling numerical and categorical variables**

we will be looking at some models

- simple linear regression
  numerical response and one predictor
- multiple linear regression
  numerical response and multiple predictors
- generalized linear models (glm)
  categorical response and predictor(s) that may be non-linear or have complicated dependence structures
  - logistic regression
  - Poisson regression

# introduction and notation

- $Y$
  **the dependent variable**
  or outcome
  or regressand
  or left-hand-side variable
  or response
  or endogeneous variable

- $X$
  **the independent variable**
  or explanatory variable
  or regressor
  or right-hand-side variable
  or treatment
  or predictor
  or covariate
  or exogeneous variable

# introduction and notation

- $Y$
  **the dependent variable**
  or outcome
  or regressand
  or left-hand-side variable
  or response
  or endogeneous variable

- $X$
  **the independent variable**
  or explanatory variable
  or regressor
  or right-hand-side variable
  or treatment
  or predictor
  or covariate
  or exogeneous variable

generally our goal is to understand how $Y$ varies as a function of $X$:

$$Y = f(X) + \text{error}$$

# why regression?

roughly, we can distinguish between three uses for regression analysis:

1. **description** - parsimonious summary of the data
2. **prediction/estimation/inference** - learn about parameters of the joint distribution of the data
3. **causal inference** - evaluate counterfactuals

# how regression?

- regression quantifies how an outcome variable $Y$ varies as a function of one or more predictor variables $X$
- the common idea: **conditioning** on $X$
- goal is to characterize $f(Y|X)$, the conditional probability distribution of $Y$ for different levels of $X$
- instead of modelling the whole conditional density of $Y$ given $X$, in regression we usually only model **the conditional mean** of $Y$ given $X$:
$$E[Y|X = x]$$
- our key goal is to approximate **the conditional expectation function** $E[Y|X]$, which summarizes how the average of $Y$ varies across all possible levels of $X$ (also called **the population regression function**)
- once we have estimated $E[Y|X]$, we can use it for **prediction** and/or **causal inference**

# linear regression

- linear regression works by assuming linear parametric form for the conditional expectation function:

$$E[Y|X] = \beta_0 + X\beta_1$$

- conditional expectation defined by only two coefficients which are estimated from the data:
  - $\beta_0$ is called the intercept or constant
  - $\beta_1$ is called the slope coefficient
- notice that the linear functional form imposes a constant slope

# interpreting the regression intercept and slope

$$E[Y|X] = \beta_0 + X\beta_1$$

when we model the regression function as a line, we can interpret the parameters of the line in appealing ways:

- **intercept**
  the average outcome among units with $X = 0$ is $\beta_0$

$$E[Y|X = 0] = \beta_0 + \beta_1(0) = \beta_0$$

- **slope**
  a one-unit change in $X$ is associated with a $\beta_1$ change in $Y$

$$\begin{aligned} E[Y|X = x + 1] - E[Y|X = x] &= (\beta_0 + \beta_1(x + 1)) - (\beta_0 + \beta_1 x) \\ &= \beta_0 + \beta_1 x + \beta_1 - \beta_0 - \beta_1 x \\ &= \beta_1 \end{aligned}$$

# linear regression

**note.** the model will not always be a good fit for the data
(even though it really wants to be)


linear regression always returns a **line** regardless of the data

# simple linear regression

**numerical response and one predictor**

## correlation

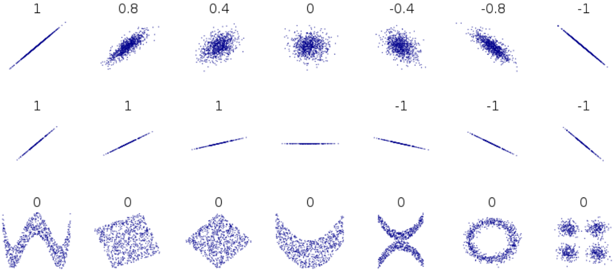
- strength of the **linear** association between two variables $(X, Y)$
- the population correlation coefficient is denoted $\rho$
- the sample correlation coefficient is denoted $r$ (or $R$)
- values between -1 (perfect negative) and +1 (perfect positive)
- value 0 indicate no linear association

# simple linear regression

**numerical response and one predictor**

## correlation

- strength of the **linear** association between two variables $(X, Y)$
- the population correlation coefficient is denoted $\rho$
- the sample correlation coefficient is denoted $r$ (or $R$)
- values between -1 (perfect negative) and +1 (perfect positive)
- value 0 indicate no linear association




from http://en.wikipedia.org/wiki/Correlation

# defining correlation

**covariance between $X$ and $Y$**

## covariance

- a generalization of variance to two random variables

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

- is not a measure of uncertainly
- a measure of how the two variables vary in relation to each other

# defining correlation

**covariance between $X$ and $Y$**

**covariance**

- a generalization of variance to two random variables

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)$$

- is not a measure of uncertainly
- a measure of how the two variables vary in relation to each other

properties of covariance:

- $\mathrm{Cov}(X, X) = \mathrm{Var}(X, X)$
- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$
- $\mathrm{Cov}(X, Y) = 0$ if $X$ and $Y$ are independent
- $\mathrm{Cov}(X, c) = 0$

# defining correlation

the Pearson correlation coefficient is given by

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \qquad -1 \leq \rho(X, Y) \leq 1$$

- $X$ and $Y$ are independent $\Rightarrow \mathrm{Cov}(X, Y) = \rho(X, Y) = 0$
- $\mathrm{Cov}(X, Y) = \rho(X, Y) = 0 \not\Rightarrow X$ and $Y$ are independent

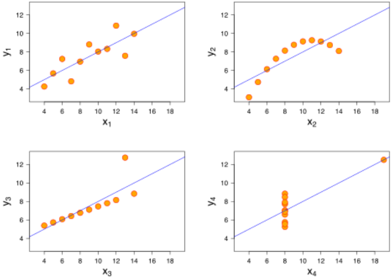
# defining correlation

the Pearson correlation coefficient is given by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \qquad -1 \leq \rho(X,Y) \leq 1$$

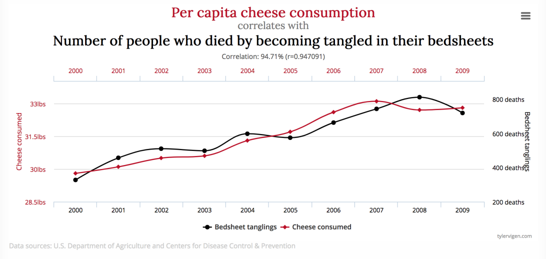
- $X$ and $Y$ are independent $\Rightarrow \mathrm{Cov}(X,Y) = \rho(X,Y) = 0$
- $\mathrm{Cov}(X,Y) = \rho(X,Y) = 0 \not\Rightarrow X$ and $Y$ are independent

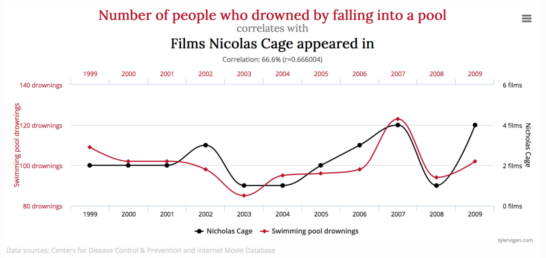
which scatterplot shows the strongest correlation?




Source: Wikipedia

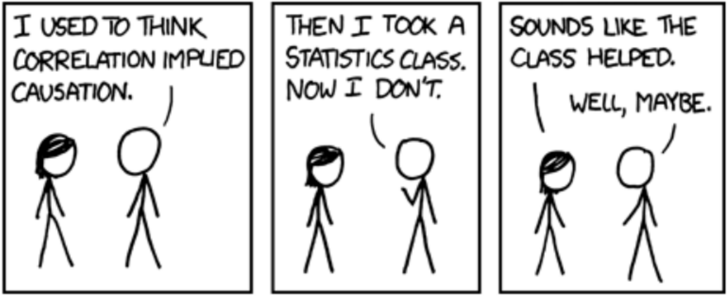
# correlation $\neq$ causation



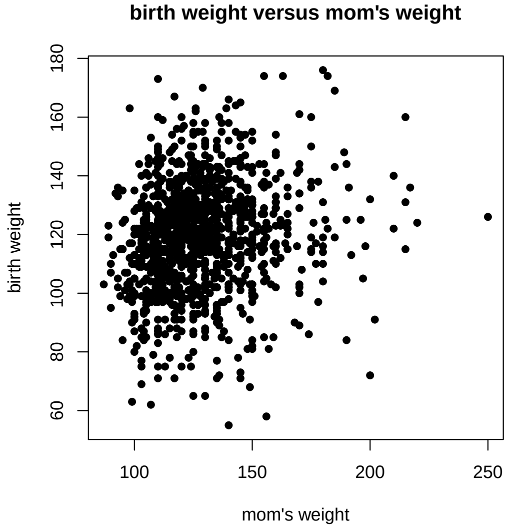

Source: `http://tylervigen.com/`

# correlation $\neq$ causation



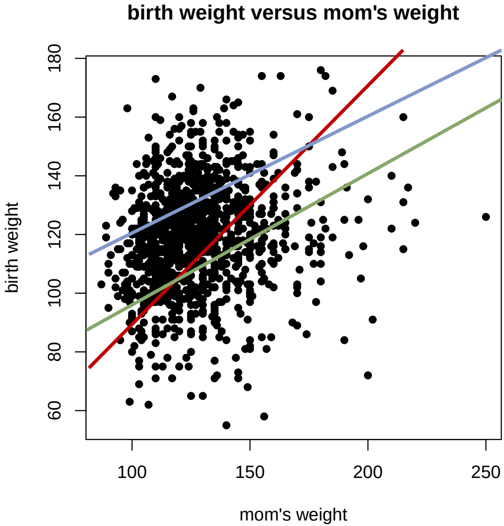
# finding the best linear fit

## example



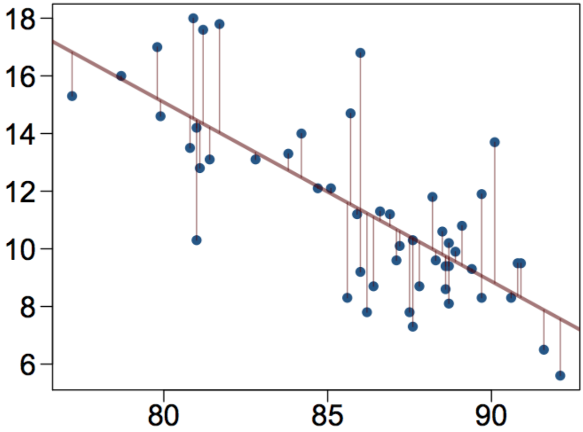
# finding the best linear fit

## example



# finding the best linear fit

## residuals




residual is the difference between the observed and predicted $y$

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

# ordinary least squares (ols)

we wish to find the line that minimises the sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

# ordinary least squares (ols)

we wish to find the line that minimises the sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

simple linear regression model has the following regression line

$$Y_i = \beta_0 + \beta_1 X_i$$

which is by ols estimated to

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

# ordinary least squares (ols)

we wish to find the line that minimises the sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

simple linear regression model has the following regression line

$$Y_i = \beta_0 + \beta_1 X_i$$

which is by ols estimated to

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

**intercept**

- parameter $\beta_0$
- point estimate $b_0$

**slope**

- parameter $\beta_1$
- point estimate $b_1$

# ordinary least squares (ols)

we wish to find the line that minimises the sum of squared residuals

$$e_1^2 + e_2^2 + \cdots + e_n^2$$

simple linear regression model has the following regression line

$$Y_i = \beta_0 + \beta_1 X_i$$

which is by ols estimated to

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

**intercept**

- parameter $\beta_0$
- point estimate $b_0$

**slope**

- parameter $\beta_1$
- point estimate $b_1$

for simple linear regression it is easy to calculate $b_0$ and $b_1$ by hand

# ordinary least squares (ols)

the slope of the regression is estimated by

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\text{sample covariance between } X \text{ and } Y}{\text{sample variance of } X}$$

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

where

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

and can be used to estimate the intercept $b_0$ by

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

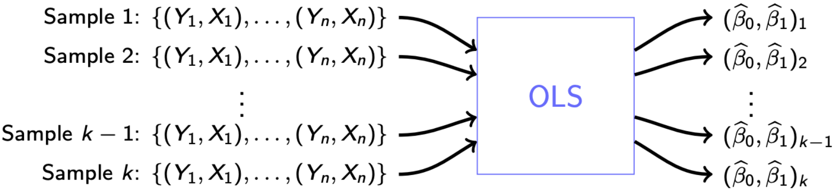
# ordinary least squares (ols)

ols is an estimator:

$$
\begin{array}{rl}
\text{Sample } 1\text{:} & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \\
\text{Sample } 2\text{:} & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \\
& \vdots \\
\text{Sample } k-1\text{:} & \{(Y_1, X_1), \ldots, (Y_n, X_n)\} \\
\text{Sample } k\text{:} & \{(Y_1, X_1), \ldots, (Y_n, X_n)\}
\end{array}
\longrightarrow \boxed{\text{OLS}} \longrightarrow
\begin{array}{l}
(\widehat{\beta}_0, \widehat{\beta}_1)_1 \\
(\widehat{\beta}_0, \widehat{\beta}_1)_2 \\
\vdots \\
(\widehat{\beta}_0, \widehat{\beta}_1)_{k-1} \\
(\widehat{\beta}_0, \widehat{\beta}_1)_k
\end{array}
$$

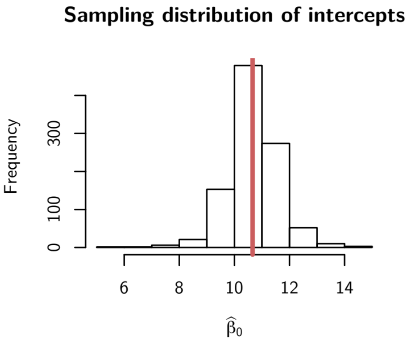
# ordinary least squares (ols)

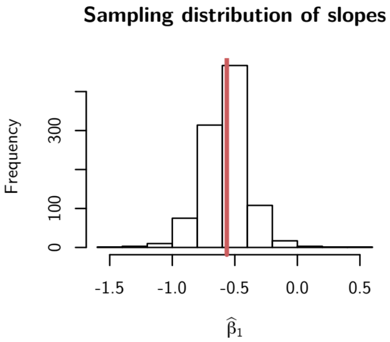
ols is an estimator:




- just like sample mean, sample difference in means, or sample variance
- it has a sampling distribution, with a sampling variance/standard error, etc.

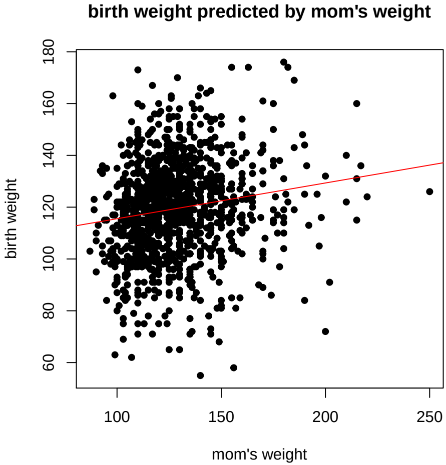
# sampling distribution of ols




the estimated slopes and intercepts vary from sample to sample, but on average the lines looks about right

# example




the estimated model is

$$\hat{y}_i = 101.75 + 0.14x_i$$

what's the interpretation?

# assessing the fit of the model

**coefficient of determination**

- denoted $r^2$ (or $R^2$) where $0 \leq r^2 \leq 1$
- what $\%$ of variability in response variable is explained by model (so optimal scenario is $r^2 = 1$)
- remainder is explained by variables not included in model

# assessing the fit of the model

**coefficient of determination**

- denoted $r^2$ (or $R^2$) where $0 \leq r^2 \leq 1$
- what % of variability in response variable is explained by model (so optimal scenario is $r^2 = 1$)
- remainder is explained by variables not included in model

**for simple linear regression**

determination coefficient is the square of correlation coefficient

# assessing the fit of the model

**coefficient of determination**

- denoted $r^2$ (or $R^2$) where $0 \leq r^2 \leq 1$
- what $\%$ of variability in response variable is explained by model (so optimal scenario is $r^2 = 1$)
- remainder is explained by variables not included in model

**for simple linear regression**
determination coefficient is the square of correlation coefficient

**example**
the correlation between birth weight and mom's weight is $r = 0.156$
then the determination coefficient is $r^2 = 0.156^2 = 0.024$

# output for example

baby weight predicted by mom's weight

$$\hat{y}_i = 101.75 + 0.14x_i$$

```
Call:
lm(formula = birth.weight ~ mom.weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-66.065 -10.943   0.333  11.048  56.075 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 101.74736    3.32162  30.632  < 2e-16 ***
mom.weight    0.13798    0.02552   5.406 7.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.12 on 1170 degrees of freedom
Multiple R-squared:  0.02437,	Adjusted R-squared:  0.02353 
F-statistic: 29.22 on 1 and 1170 DF,  p-value: 7.819e-08
```

# output for example

baby weight predicted by mom's weight

$$\hat{y}_i = 101.75 + 0.14x_i$$

```
Call:
lm(formula = birth.weight ~ mom.weight)

Residuals:
    Min      1Q  Median      3Q     Max
-66.065 -10.943   0.333  11.048  56.075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.74736    3.32162  30.632  < 2e-16 ***
mom.weight    0.13798    0.02552   5.406 7.82e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.12 on 1170 degrees of freedom
Multiple R-squared:  0.02437,	Adjusted R-squared:  0.02353
F-statistic: 29.22 on 1 and 1170 DF,  p-value: 7.819e-08
```

# output for example

baby weight predicted by mom's weight

$$\hat{y}_i = 101.75 + 0.14x_i$$

```
Call:
lm(formula = birth.weight ~ mom.weight)

Residuals:
    Min      1Q  Median      3Q     Max
-66.065 -10.943   0.333  11.048  56.075

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.74736    3.32162  30.632  < 2e-16 ***
mom.weight    0.13798    0.02552   5.406 7.82e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.12 on 1170 degrees of freedom
Multiple R-squared:  0.02437,	Adjusted R-squared:  0.02353
F-statistic: 29.22 on 1 and 1170 DF,  p-value: 7.819e-08
```

# output for example

baby weight predicted by mom's weight

$$\hat{y}_i = 101.75 + 0.14x_i$$

```
Call:
lm(formula = birth.weight ~ mom.weight)

Residuals:
    Min      1Q  Median      3Q     Max 
-66.065 -10.943   0.333  11.048  56.075 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 101.74736    3.32162  30.632  < 2e-16 ***
mom.weight    0.13798    0.02552   5.406 7.82e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.12 on 1170 degrees of freedom
Multiple R-squared:  0.02437,	Adjusted R-squared:  0.02353 
F-statistic: 29.22 on 1 and 1170 DF,  p-value: 7.819e-08
```

# coefficient of determination

- $$R^2 = 1 - \frac{RSS}{TSS} \quad RSS = \sum_{i=1}^{n} e_i^2 \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- Caveat: when a covariate is added, $R^2$ increases

# coefficient of determination

- $$R^2 = 1 - \frac{RSS}{TSS} \quad RSS = \sum_{i=1}^{n} e_i^2 \quad TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$
- Caveat: when a covariate is added, $R^2$ increases
- Adjusted $R^2$, or $R^2_{adj}$

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- $p$ - number of covariates in the model

# Part II

# multiple linear regression

same underlying idea as before but with multiple predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{Np}$$

which is by ols estimated to (requires matrix algebra)

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{np}$$

# multiple linear regression

same underlying idea as before but with multiple predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{Np}$$

which is by ols estimated to (requires matrix algebra)

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{np}$$

```
Call:
lm(formula = birth.weight ~ gestation + mom.smokes)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.553 -10.855  -0.178  10.013  50.495 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -13.64855    8.69303  -1.570    0.117    
gestation     0.48809    0.03095  15.771   <2e-16 ***
mom.smokes   -8.17175    0.96916  -8.432   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.15 on 1169 degrees of freedom
Multiple R-squared:  0.2259,	Adjusted R-squared:  0.2246 
F-statistic: 170.6 on 2 and 1169 DF,  p-value: < 2.2e-16
```

for interpretation we must use **’all else held constant’**

# conditions for inference

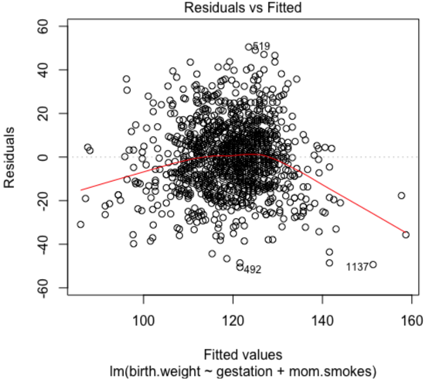
model assumptions must be met:

- linearity
- constant residual variance (homoscedasticity and no autocorrelation)
- approximately normally distributed residuals

# conditions for inference

model assumptions must be met:

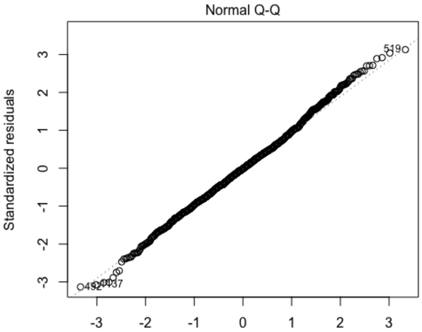
- **linearity**
- **constant residual variance (homoscedasticity and no autocorrelation)**
- approximately normally distributed residuals



# conditions for inference

model assumptions must be met:

- linearity
- constant residual variance (homoscedasticity and no autocorrelation)
- **approximately normally distributed residuals**



# conditions for inference

model assumptions must be met:

- linearity
- constant residual variance (homoscedasticity and no autocorrelation)
- approximately normally distributed residuals

**and most importantly**

- the response is continuous variable that is normally distributed

*...but what if it isn't?*

# generalized linear models (glm)

a glm has the following three components:

1. a probability distribution describing the response variable $Y$ that should belong to **the exponential family**:
   - normal
   - binomial
   - Poisson

     $\vdots$
2. a linear function of the regressors, called **linear predictor**

$$\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

   on which the expected value $\mu_i$ of $Y_i$ depend

3. a invertible **link function** which transforms the expectation of the response to the linear predictor

$$g(\mu_i) = \eta_i \quad \text{or} \quad g^{-1}(\eta_i) = \mu_i$$

# logistic regression

- assume a binomial distribution produced the outcome variable
- want to model $\mu = p$ (probability of success) given set of predictors

the logistic model is specified when a link function connects $\eta$ to $p$

# logistic regression

- assume a binomial distribution produced the outcome variable
- want to model $\mu = p$ (probability of success) given set of predictors

the logistic model is specified when a link function connects $\eta$ to $p$

the most commonly used is the logit function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad \text{for } 0 \leq p \leq 1$$

the logit function takes a value between 0 and 1 and maps it to a value between $-\infty$ and $\infty$

the inverse of the logit function:

$$g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} = \frac{1}{1+\exp(-x)}$$

the inverse logit function takes a value between $-\infty$ and $\infty$ and maps it to a value between 0 and 1

# logistic regression

the three glm criteria give us

1. $y_i \sim \text{Binomial}(p_i, n)$
2. $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
3. $\text{logit}(p_i) = \eta_i$

which gives us

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

and

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}$$

# logistic regression

the three glm criteria give us

1. $y_i \sim \text{Binomial}(p_i, n)$
2. $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
3. $\text{logit}(p_i) = \eta_i$

which gives us

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

and

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}$$

# logistic regression

the three glm criteria give us

1. $y_i \sim \text{Binomial}(p_i, n)$
2. $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
3. $\text{logit}(p_i) = \eta_i$

which gives us

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

and

$$p_i = \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}\right)}$$

# logistic regression

## odds

odds are another way to quantify the probability of an event

# logistic regression

## odds

odds are another way to quantify the probability of an event

for some event $E$

$$\mathrm{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

# logistic regression
## odds

odds are another way to quantify the probability of an event

for some event $E$

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

if we are told the odds of $E$ are $x$ to $y$, then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

$$\implies P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

# logistic regression
## odds

odds are another way to quantify the probability of an event

for some event $E$

$$\text{odds}(E) = \frac{P(E)}{P(E^c)} = \frac{P(E)}{1 - P(E)}$$

if we are told the odds of $E$ are $x$ to $y$, then

$$\text{odds}(E) = \frac{x}{y} = \frac{x/(x+y)}{y/(x+y)}$$

$$\implies P(E) = x/(x+y), \quad P(E^c) = y/(x+y)$$

useful when interpreting coefficient estimates of a logistic regression

# logistic regression

**example. lab 5**

we want to create a spam filter based on

- 3921 observations/emails
- properties of the emails (more details during class)

# logistic regression

**example. lab 5**

we want to create a spam filter based on

- 3921 observations/emails
- properties of the emails (more details during class)

simple model: use binary predictor 'winner'

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_{1i} \times \text{winner}$$

which is estimated to

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

# logistic regression

**example. lab 5**

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

# logistic regression
## example. lab 5

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

the odds of an email being spam if 'winner' = 0:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \exp(-2.31) = 0.10$$

# logistic regression

**example. lab 5**

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

the odds of an email being spam if 'winner' = 0:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \exp(-2.31) = 0.10$$

the probability of an email being spam if 'winner' = 0:

$$\hat{p}_i = \frac{0.1}{1.1} = 0.09$$

# logistic regression

**example. lab 5**

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

the odds of an email being spam if 'winner' = 0:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \exp(-2.31) = 0.10$$

the probability of an email being spam if 'winner' = 0:

$$\hat{p}_i = \frac{0.1}{1.1} = 0.09$$

what are the odds and probability if 'winner' = 1?

# logistic regression

**example. lab 5**

$$\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = -2.31 + 1.53 \times \text{winner}$$

the odds of an email being spam if 'winner' = 0:

$$\frac{\hat{p}_i}{1-\hat{p}_i} = \exp(-2.31) = 0.10$$

the probability of an email being spam if 'winner' = 0:

$$\hat{p}_i = \frac{0.1}{1.1} = 0.09$$

what are the odds and probability if 'winner' = 1?

answer: odds is 0.45 and probability is 0.31

# output for spam filter example

```
Call:
glm(formula = spam ~ winner, family = binomial, data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8657  -0.4342  -0.4342  -0.4342   2.1947

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31405    0.05627 -41.121  < 2e-16 ***
winneryes    1.52559    0.27549   5.538 3.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2412.7  on 3919  degrees of freedom
AIC: 2416.7

Number of Fisher Scoring iterations: 5
```

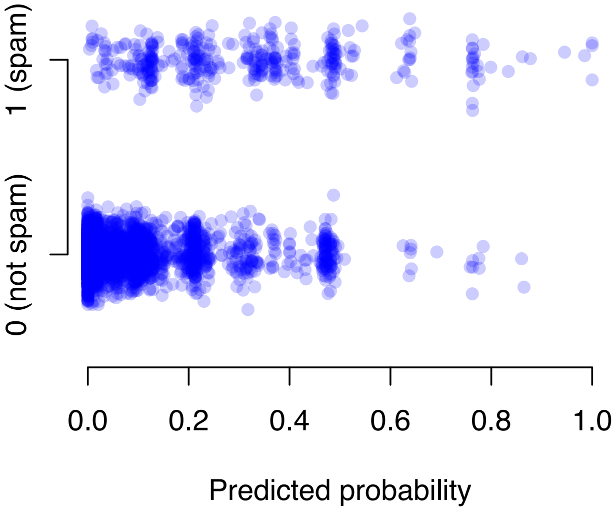
# output for spam filter example

```
Call:
glm(formula = spam ~ winner, family = binomial, data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8657  -0.4342  -0.4342  -0.4342   2.1947

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31405    0.05627 -41.121  < 2e-16 ***
winneryes    1.52559    0.27549   5.538 3.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2412.7  on 3919  degrees of freedom
AIC: 2416.7

Number of Fisher Scoring iterations: 5
```

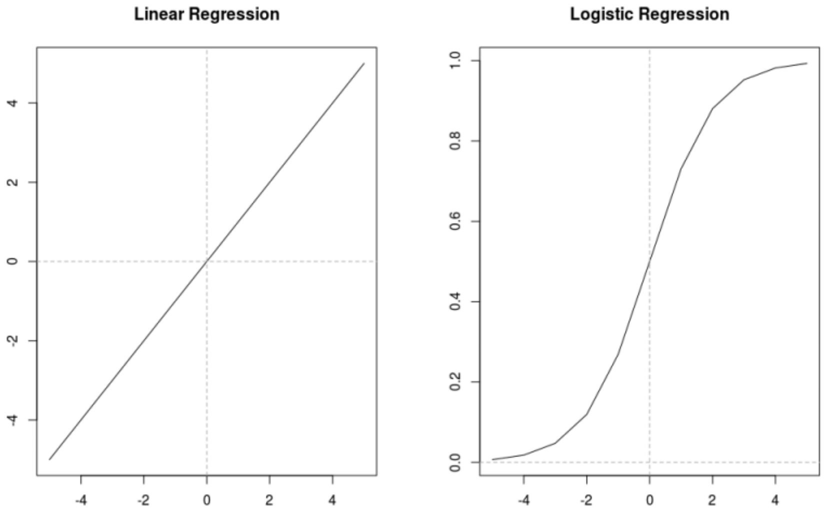
# output for spam filter example

```
Call:
glm(formula = spam ~ winner, family = binomial, data = email)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8657  -0.4342  -0.4342  -0.4342   2.1947

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31405    0.05627 -41.121  < 2e-16 ***
winneryes    1.52559    0.27549   5.538 3.06e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2412.7  on 3919  degrees of freedom
AIC: 2416.7

Number of Fisher Scoring iterations: 5
```

# output for spam filter example



# linear versus logistic regression



# Poisson regression

suitable for when the response is count data

the three glm criteria give us

1. $y_i \sim \text{Poisson}(\mu_i)$
2. $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$
3. $\log(\mu_i) = \eta_i$

which gives us

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

and

$$\mu_i = \exp\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{ki}\right)$$

# Poisson regression

**example in r**

$y_i$ is number of spam emails received, $y_i \sim \text{Poisson}(\mu_i)$
predict $y_i$ using total nr of emails received

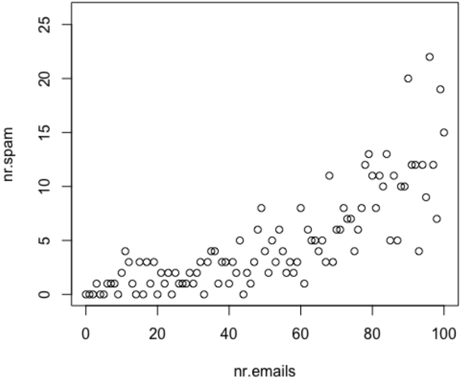
$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{nr.emails}$$

# Poisson regression

**example in r**

$y_i$ is number of spam emails received, $y_i \sim \text{Poisson}(\mu_i)$
predict $y_i$ using total nr of emails received

$$\log(\mu_i) = \beta_0 + \beta_1 \times \text{nr.emails}$$



# Poisson regression

## example in r

```
Call:
glm(formula = nr.spam ~ nr.emails, family = poisson, data = MyData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.75498  -0.98292  -0.07578   0.53401   2.35312

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.253771   0.143496  -1.768    0.077 .
nr.emails    0.029698   0.001889  15.724   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 424.24  on 100  degrees of freedom
Residual deviance: 123.01  on  99  degrees of freedom
AIC: 417.35

Number of Fisher Scoring iterations: 5
```

# Poisson regression

## example in r

```
Call:
glm(formula = nr.spam ~ nr.emails, family = poisson, data = MyData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.75498  -0.98292  -0.07578   0.53401   2.35312

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.253771   0.143496  -1.768    0.077 .
nr.emails    0.029698   0.001889  15.724   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 424.24  on 100  degrees of freedom
Residual deviance: 123.01  on  99  degrees of freedom
AIC: 417.35

Number of Fisher Scoring iterations: 5
```

# Poisson regression

## example in r

```
Call:
glm(formula = nr.spam ~ nr.emails, family = poisson, data = MyData)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.75498  -0.98292  -0.07578   0.53401   2.35312

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.253771   0.143496  -1.768    0.077 .
nr.emails    0.029698   0.001889  15.724   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 424.24  on 100  degrees of freedom
Residual deviance: 123.01  on  99  degrees of freedom
AIC: 417.35

Number of Fisher Scoring iterations: 5
```

# Poisson regression

**example in r**

estimated model is $\hat{\mu}_i = \exp\left(-0.25 + 0.03 \times \text{nr.emails}\right)$

# Poisson regression

### example in r

estimated model is $\hat{\mu}_i = \exp\left(-0.25 + 0.03 \times \text{nr.emails}\right)$

assume you receive 30 emails, how many are expected to be spam?

$$\hat{\mu}_i = \exp\left(-0.25 + 0.03 \times 30\right) = 1.91 \approx 2$$

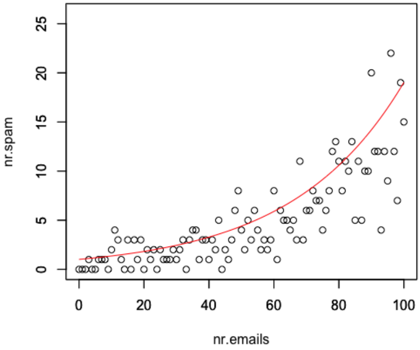
# Poisson regression

**example in r**

estimated model is $\hat{\mu}_i = \exp(-0.25 + 0.03 \times \text{nr.emails})$

assume you receive 30 emails, how many are expected to be spam?

$$\hat{\mu}_i = \exp(-0.25 + 0.03 \times 30) = 1.91 \approx 2$$



# Poisson regression

**example in r**

- in Poisson distribution, $E(X) = Var(X) = \mu$
- problem of overdispersion $E(X) < Var(X)$

# Poisson regression

**example in r**

- in Poisson distribution, $E(X) = Var(X) = \mu$
- problem of overdispersion $E(X) < Var(X)$
- `family=quasipoisson`

# Poisson regression
**example in r**

- in Poisson distribution, $E(X) = Var(X) = \mu$
- problem of overdispersion $E(X) < Var(X)$
- `family=quasipoisson`
- negative binomial distribution

# different link functions

| link name | $\eta_i = g(\mu_i)$ | $\mu_i = g^{-1}(\eta_i)$ |
|---|---|---|
| identity | $\mu_i$ | $\eta_i$ |
| log | $\log(\mu_i)$ | $\exp(\eta_i)$ |
| logit | $\log\left(\frac{\mu_i}{1-\mu_i}\right)$ | $\frac{1}{1+\exp(-\eta_i)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

# different link functions

| link name | $\eta_i = g(\mu_i)$ | $\mu_i = g^{-1}(\eta_i)$ |
|---|---|---|
| identity | $\mu_i$ | $\eta_i$ |
| log | $\log(\mu_i)$ | $\exp(\eta_i)$ |
| logit | $\log\left(\frac{\mu_i}{1-\mu_i}\right)$ | $\frac{1}{1+\exp(-\eta_i)}$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

| family | canonical link | range of $Y_i$ |
|---|---|---|
| Gaussian (normal) | identity | $(-\infty, \infty)$ |
| binomial | logit | $\{0,1\}$ |
| Poisson | log | $0, 1, 2, \ldots$ |
| $\vdots$ | $\vdots$ | $\vdots$ |

# estimation of glm

- glms are estimated via **maximum likelihood estimation** (mle)
- **most likely** values of parameters **given the data** we observed
- applied to linear regression $\implies$ ols=mle

# estimation of glm

- glms are estimated via **maximum likelihood estimation** (mle)
- **most likely** values of parameters **given the data** we observed
- applied to linear regression $\implies$ ols=mle

## model fit

conditions for inference are the same as for linear regression
(recall that linear regression is just a special case of glm)

goodness of fit measures:

- Pearson chi-square statistic
- deviance: $-2(\log L_m - \log L_s)$
  (`anova(model1,test=''chisq'')`)
- likelihood ratio tests (`epicalc::lrtest(model1, model2)`)

# estimation of glm

- glms are estimated via **maximum likelihood estimation** (mle)
- **most likely** values of parameters **given the data** we observed
- applied to linear regression $\implies$ ols=mle

## model fit

conditions for inference are the same as for linear regression
(recall that linear regression is just a special case of glm)

goodness of fit measures:

- Pearson chi-square statistic
- deviance: $-2(\log L_m - \log L_s)$

  (`anova(model1,test=''chisq'')`)

```
        Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     3920     2437.2
winner   1   24.506      3919     2412.7 7.41e-07 ***
```

- likelihood ratio tests (`epicalc::lrtest(model1, model2)`)

# estimation of glm

- glms are estimated via **maximum likelihood estimation** (mle)
- **most likely** values of parameters **given the data** we observed
- applied to linear regression $\implies$ ols=mle

**model fit**
conditions for inference are the same as for linear regression
(recall that linear regression is just a special case of glm)

goodness of fit measures:

- Pearson chi-square statistic
- deviance: $-2(\log L_m - \log L_s)$
  (`anova(model1,test=''chisq'')`)
- likelihood ratio tests (`epicalc::lrtest(model1, model2)`)
- Akaike Information Criterion (AIC) – Lecture 9

# reading

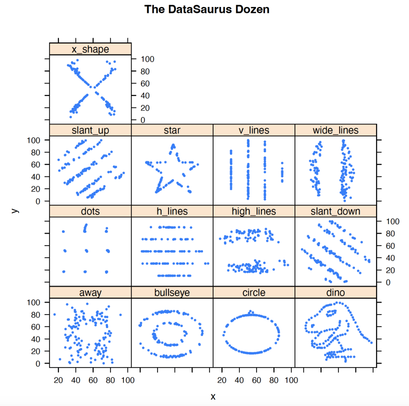
Agresti A., 2018, Statistical Methods for the Social Sciences, Fifth Edition, **Chapters 9, 11, 14.4, 15.1**

**link to the book via Manchester library**

# reading

Agresti A., 2018, Statistical Methods for the Social Sciences, Fifth Edition, **Chapters 9, 11, 14.4, 15.1**

**link to the book via Manchester library**

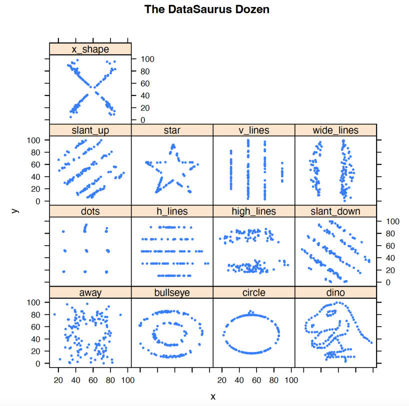



Source: Wikipedia

# reading

Agresti A., 2018, Statistical Methods for the Social Sciences, Fifth Edition, **Chapters 9, 11, 14.4, 15.1**

**link to the book via Manchester library**




Source: Wikipedia

$$r = 0. - 0.06$$