PERSONAL - Applied Regression

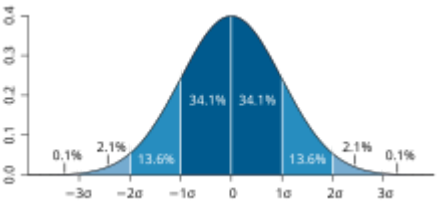
- Recap: Basic Statistics
- Introduction
- Simple Linear Regression
- Recap: Matrix Algebra
  - Operations and Special Types
- Simple Regression (Matrix)
- Multiple Regression
- Specification
- Model Diagnostics
- Lack of Fit
- Model Selection
- Nonlinear Regression
- Time Series Models
- Logistic Regression
- Poisson Regression
- Linear Mixed Effects Models
- Statistical Learning (Machine Learning)
  - Prediction Methods
  - Statistical Decision Theory
  - Categorical Data

Recap: Basic Statistics

- **mean (expected value)**: the mean of the possible values a random variable can take
  - **notation**: $E(X) = \mu$
    - **finitely many outcomes**: $E(X) = x_1p_1 + ... + x_np_n$
    - **countably infinitely many outcomes**: $E(X) = \sum_{i=1}^{\infty} x_ip_i$
    - **random variables with density**: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$
  - **properties**:
    - $E(c) = c$
    - $E(aX + b) = aE(x) + b$
    - $E(a_1X_1 + ... + a_nX_n) = a_1E(X_1) + ... + a_nE(X_n)$

- $E(X_1 \cdot ... \cdot X_n) = E(X_1) \cdot ... \cdot E(X_n)$ for independent (uncorrelated) X_i
- $E(g(X)) = \int_{\mathbb{R}} g(x)f(x)dx$

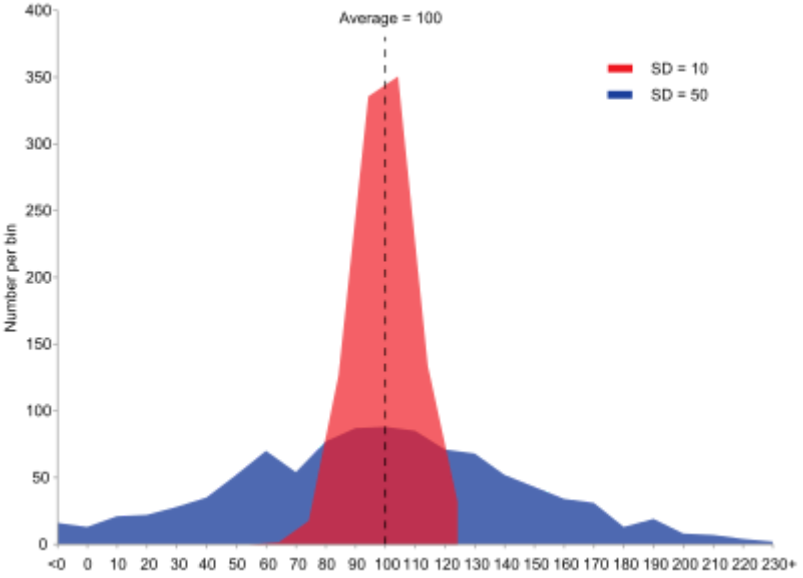
- **standard deviation**: a measure of the amount of variation of the values of a variable about its mean



  - **definition**: $\sigma = \sqrt{Var(X)}$

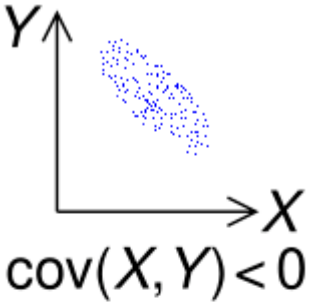
- **standard error**: the standard deviation of its sampling distribution or an *estimate* of that standard deviation
  - **definition**: $se = \frac{\sigma}{\sqrt{n}}$ for n observations

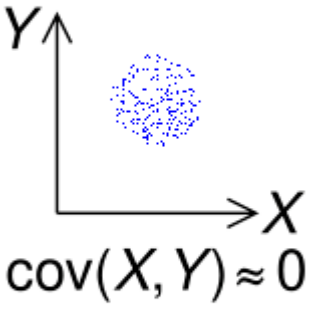
- **variance**: a measure of how far a set of numbers is spread out from their average value

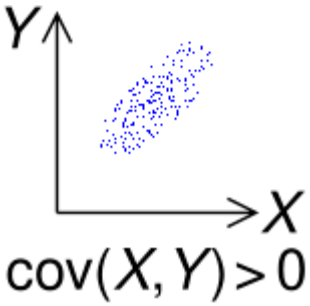


  - **notation**: $Var(X) = \sigma^2$
  - **definition**: $Var(X) = E(X^2) - E(X)^2$
  - **properties**:
    - $Var(c) = 0$
    - $Var(X + a) = Var(X)$
    - $Var(aX) = a^2Var(X)$
    - $Var(aX \pm bY) = a^2Var(X) + b^2Var(Y) \pm 2ab\ Cov(X, Y)$
    - $Var(X_1 + ... + X_n) = Var(X_1) + ... + Var(X_n)$ for independent (uncorrelated) X_i

- **covariance**: a measure of the joint variability of two random variables







  - **definition**: $Cov(X,Y) = E(XY) - E(X)E(Y)$
  - **positively correlated variables**: $Cov(X,Y) > 0$
  - **negatively correlated variables**: $Cov(X,Y) < 0$
  - **uncorrelated variables**: $Cov(X,Y) = 0$ (but not the other way around!)
    - **independent variables**: $Cov(X,Y) = 0$
  - **properties**:
    - $Cov(X,X) = Var(X)$
    - $Cov(X,Y) = Cov(Y,X)$
    - $Cov(X,c) = 0$
    - $Cov(aX,bY) = ab\ Cov(X,Y)$
    - $Cov(X+a,Y+b) = Cov(X,Y)$
    - $Cov(aX+bY,cW+dV) = acCov(X,W) + adCov(X,V) + bcCov(Y,W) + bdCov(Y,V)$
- **correlation**: any statistical relationship, whether causal or not, between two random variables
  - **definition**: $Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

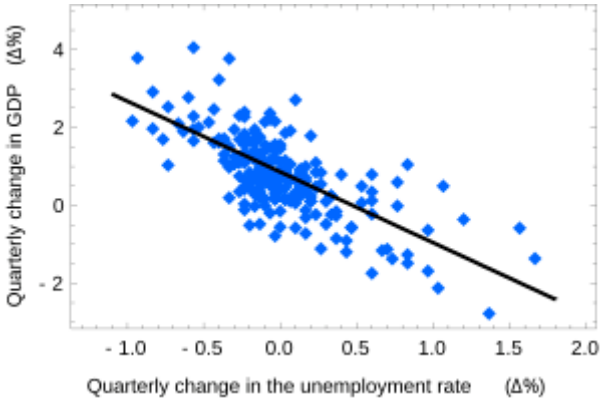
Introduction

- **regression**: creates a functional relationship between a **response (dependent) variable** and a set of **explanatory (predictor) variables (covariates)**
  - **regression model**: which explanatory variables have an effect on the response?
- **deterministic relationship**: a certain input will *always* lead to the same result
- **parameter**: an unknown constant, most likely to be estimated by collecting and using data
- **empirical model**: any kind of model based on **empirical observations** rather than on **mathematically describable** (theory-based) relationships of the system modelled
- **controlled experiment**: one where the experimenter can set the values of the explanatory variable(s)
- **line definition (linear model)**: $y = \beta_0 + \beta_1 x \ (+\epsilon)$
  - β_0, β_1: constants (parameters)
    - **intercept** β_0: y when $x = 0$
    - **slope** β_1: change in y if x is increased by 1 unit
  - ϵ: random disturbance (error)
    - $\beta_0 + \beta_1 x$: *deterministic*
    - ϵ: *random*, models **variability** in measurements around the **regression line**
  - *linear* in β_0 and β_1
  - **for each experiment**: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
    - **input and result**: (x_i, y_i)
    - β_0, β_1 remain *constant*
    - x_i, ϵ_i vary per experiment $i = 1, 2, ..., n$
      - mean $E(\epsilon_i) = 0$
      - variance $Var(\epsilon_i) = \sigma^2$
      - ϵ_i, ϵ_j independent *random* variables for $i \neq j$
      - x_i *deterministic* (i.e. the input data is clearly and certainly defined; it can also be noisy, in which case x_i is not deterministic)
    - $\implies$ y_i *random* variable; y_i, y_j independent for $i \neq j$
      - mean $E(y_i) = E(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i + \underbrace{E(\epsilon_i)}_{0} = \beta_0 + \beta_1 x_i$
      - variance $Var(y_i) = \sigma^2$
        - unexplained variability σ
  - **general**: $y = \mu + \epsilon$

        - **deterministic component** $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$
            - **explanatory variables** $x_1, ..., x_p$ (assume fixed, measured without error)
            - $\beta_i, i = 1, 2, ..., p$: change in μ when changing x_i by one unit while keeping *all* other explanatory variables the same
            - $E(y) = \mu, Var(y) = \sigma^2$
    - **linearity**: the derivatives of μ with respect to the parameters β_i do *not* depend on the variables
- **notation**: x_{ij} for the i-th unit (i.e. row in a table) and the j-th explanatory variable (i.e. column in a table) (R: `table[i,j]`)
- **dependent variable**: depends on an **independent** variable

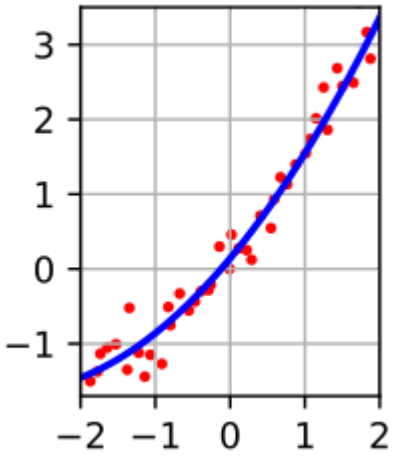
Simple Linear Regression

- simple linear regression model: $y = \mu + \epsilon$



    - **mean** $E(y) = \mu = \beta_0 + \beta_1 x$
    - one **predictor (regressor)** variable x
    - one **response** variable y
    - **random error** ϵ
    - for n pairs of observations (x_i, y_i): $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, ..., n$
        - x_i *not* random (can be selected by experimenter)
        - $\epsilon_i \sim N(0, \sigma^2)$
            - $y_i \sim N(\mu_i, \sigma^2)$, where $\mu_i = \beta_0 + \beta_1 x_i$
        - $E(\epsilon_i) = 0$
            - $E(y_i) = \mu_i = \beta_0 + \beta_1 x_i$
        - $Var(\epsilon_i) = \sigma^2$
            - $Var(y_i) = \sigma^2$
        - $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$
            - any two observations y_i, y_j are *independent* for $i \neq j$
    - **goal**: estimate $\beta_0, \beta_1, \sigma^2$ from available data (x_i, y_i)

    - zero slope $\implies$ absence of linear association
- **unbiased parameter estimate**: $E(\hat{\theta}) = \theta$
    - **biased parameter estimate**: $E(\hat{\theta}) \neq \theta$
- [least squares estimation (LSE)]: a mathematical procedure for *finding the best-fitting curve to a given set of points* by *minimizing the sum of the squares* of the offsets ("the residuals") of the points from the curve

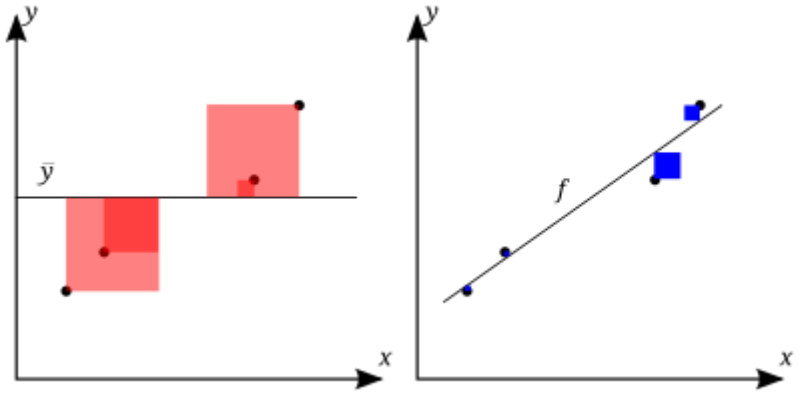


    - **goal**: minimize $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ (fitted value)
        - **LSE** $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{s_{xx}}$
            - $s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
            - $s_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})$
            - **(!)** reordering using $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$:
                - $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$
                - $s_{xx} = \sum_{i=1}^{n} x_i(x_i - \overline{x})$
        - **LSE** $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$
        - **LSE** $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
            - short: $s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$
            - **residual** $e_i = y_i - \hat{y}_i$
            - **degree of freedom**: number of independent observations (n) minus the number of estimated parameters (here 2, β_0 and β_1)
        - **sample mean** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
        - **result mean** $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
            - $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$
        - $E(\hat{\beta}_1) = \beta_1$
        - $E(\hat{\beta}_0) = \beta_0$
        - $E(\overline{y}) = \beta_0 + \beta_1 \overline{x}$

    - $E(s^2) = \sigma^2$
    - $Var(\hat{\beta}_1) = \frac{\sigma^2}{s_{xx}}$
    - $Var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}} \right)$
    - $se(\hat{\beta}_1) = \frac{s}{\sqrt{s_{xx}}}$
- **maximum likelihood estimation (MLE)**: a method of *estimating the parameters* of an assumed probability distribution by *maximizing a likelihood function*
    - $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ (*biased*!)
- **null hypothesis testing**: a method of statistical inference used to decide whether the data sufficiently supports a particular hypothesis
    - t**-test**: a statistical test used to test whether the difference between the response of two groups is statistically significant or not (here: two-sided)
        - $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$ ($\leq$ or $>$)
        - $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-2}$
        - α usually 0.05
            - **quantile approach**: reject H_0 if $|T| > t_{n-2,1-\alpha/2}$
            - **probability approach**: reject H_0 if p-value is *less* than α
                - p-value: the probability of obtaining test results at least as extreme as the result actually observed, under the assumption that the null hypothesis is correct
                    - **R**: `2 * pt(abs(tval), df, lower.tail = FALSE)`
                - the lower the p-value, the more far-fetched the null hypothesis is
- **confidence interval**: an interval which is expected to typically contain the parameter being estimated
    - $100(1-\alpha)\%$ **confidence interval for** β_1: $\hat{\beta}_1 \pm t_{n-2,1-\alpha/2} \cdot se(\hat{\beta}_1)$
        - **general**: $\text{Estimate} \pm (\text{t value})(\text{standard error of estimate})$
- **prediction of a new point**: $y_p = \hat{\beta}_0 + \hat{\beta}_1 x + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$
- **analysis of variance (ANOVA)**: a collection of statistical models and their associated estimation procedures used to analyze the differences between groups
    - $SST = SSR + SSE$

Source	d.f.	SS (Sum of Squares)	MS (Mean Square)	F
Regression	1	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	$MSR = SSR$	$\frac{MSR}{MSE}$
Residual (Error)	$n-2$	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$MSE = s^2$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$		

- [F-test]: any statistical test used to compare the variances of two samples or the ratio of variances between multiple samples
  - $H_0 : \beta_1 = 0$ vs. $H_A : \beta_1 \neq 0$
  - $T \sim t_v \implies T^2 \sim F_{1,v}$
  - **quantile approach**: reject H_0 if $F > F_{1,n-2,1-\alpha}$ (`qf(1 - alpha, 1, n - 2)`)
  - **probability approach**: reject H_0 if $P(f > F) < \alpha$ where $f \sim F_{1,n-2}$
- [coefficient of determination]: the proportion of the variation in the dependent variable that is predictable from the independent variable(s)



  - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
    - **interpretation**: $100R^2\%$ of the variation in y can be explained by x
  - $0 \leq R^2 \leq 1$
    - the better the linear regression (right) fits the data in comparison to the simple average (left), the closer the value of R^2 is to 1
- [Pearson correlation]: a correlation coefficient that measures **linear (!)** correlation between two variables x, y
  - $r = \text{sign}(\hat{\beta}_1)\sqrt{R^2}$
  - **in R**...
    - `cov(x,y)` returns $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$
    - `var(x)` returns $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$
    - `cor(x,y)` returns $\frac{Cov(x,y)}{sd(x)sd(y)}$
    - `sd(x)` returns `sqrt(var(x))`
- **diagnostics**: $y \sim N(\beta_0 + \beta_1 x, \sigma^2)$
  - **independence**: told by investigator
  - **linearity**: plot y against x
  - **constant variance**: plot $y - \hat{y}$ against $\hat{y}$
  - **normal distribution**: plot $y - \hat{y}$ against normal quantiles

Recap: Matrix Algebra

- **matrix**: a rectangular array of numbers
  - **formally**: $A \in p \times q$ (matrix A with p rows and q columns)
    - $A = (a_{ij})$, where a_{ij} is the entry in row i and column j
  - **square matrix**: same number of rows and columns ($p = q$)
    - $det(AB) = det(A)det(B)$ (also written as $|AB| = |A||B|$)
  - **identity matrix** I: square matrix with **ones** in the diagonal and zeros everywhere else
  - **zero matrix** O: matrix of all **zeros**
  - **diagonal (square) matrix**: all entries outside the diagonal are zero
- **(column) vector**: a matrix consisting of a single column
  - **formally**: $\boldsymbol{x} \in p \times 1$
    - $\boldsymbol{x} = (x_i)$
  - **elements**: $x_1, ..., x_p$
  - **unit vector** $\mathbf{1}$: vector with all elements equal to **one**
  - **zero vector** $\mathbf{0}$: vector with all elements equal to **zero**

Operations and Special Types

- **matrix / vector addition**: element-wise, **same** dimensions
- **matrix / vector multiplication**: for $A \in p \times q, \quad B \in q \times t$, go through each row in the first matrix and multiply and add the elements with the elements of each column in the second matrix; that's one complete row in the result matrix
  - **formally**: $C = AB = (c_{ij}) \in p \times t$ with $c_{ij} = \sum_{r=1}^{q} a_{ir} b_{rj}$
  - $(AB)C = A(BC)$
  - $(A + B)C = AC + BC$
  - $A(B + C) = AB + AC$
- **matrix transposition**: interchange rows and columns
  - **formally**: $A' = (a_{ji})$ with $A \in q \times p$
  - **symmetric matrix**: $A = A'$
  - $(A + B)' = A' + B'$
  - $(A')' = A$
  - $(cA)' = cA'$
  - $(AB)' = B'A'$ for $A \in m \times n, B \in n \times p$
- **(inner) vector product**: multiply element-wise, then add all together $\rightarrow$ scalar

    - **formally**: $\boldsymbol{x'y} = \sum_{i=1}^{p} x_i y_i$
    - **orthogonal vectors**: inner product 0
    - **euclidian norm (length)**: $||\boldsymbol{x}|| = \sqrt{\boldsymbol{x'x}}$
- **set of linearly dependent vectors**: there exist scalars c_i, not all simultaneously zero, such that $c_1\boldsymbol{x_1} + ... + c_k\boldsymbol{x_k} = 0$
    - **(!)** at least one vector can be written as a linear combination of the remaining ones (for example, a column in a matrix is the summation of two other columns)
    - **linearly independent**: otherwise
- **matrix rank**: largest number of linearly independent columns (or rows)
    - **nonsingular matrix**: square matrix with rank equal to row / column number
        - **formally**: $A \in m \times m, \;\; rank(A) = m$
- **matrix inverse**: $AA^{-1} = A^{-1}A = I$
    - $ABB^{-1}A^{-1} = I$
    - $(A^{-1})' = (A')^{-1}$
    - $(\lambda A)^{-1} = \frac{1}{\lambda}A^{-1}$
    - **for nonsingular matrices**: $(AB)^{-1} = B^{-1}A^{-1}$
- **orthogonal (square) matrix**: $AA' = A'A = I$
    - $A' = A^{-1}$
    - the rows (columns) are **mutually orthogonal**
    - the **length** of the rows (columns) is **one**
    - $det(A) = \pm 1$
- **trace of a (square) matrix**: the sum of its diagonal elements
    - **formally**: $tr(A) = \sum_{i=1}^{m} a_{ii}$
    - $tr(A) = tr(A')$
    - $tr(A + B) = tr(A) + tr(B)$
    - $tr(CDE) = tr(ECD) = tr(DEC)$ for conformable matrices C, D, E (matrices s.t. products are defined)
    - $tr(c) = c$
    - **bonus**: $E(tr(\cdot)) = tr(E(\cdot))$
- **idempotent (square) matrix**: $AA = A$
    - $det(A) = 0$ or 1
    - $rank(A) = tr(A)$

Simple Regression (Matrix)

- **simple regression (matrix approach)**: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$
  - $\boldsymbol{y}, \boldsymbol{\epsilon}$ are $(n \times 1)$ **random** vectors
  - $\boldsymbol{X}$ is a $(n \times 2)$ matrix (first col. ones, second column x_i
  - **LSE** $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$
  - **fitted value vector** $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$
  - **residual vector** $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$
  - **LSE** $s^2 = \frac{1}{n-2}\boldsymbol{e'e}$
- **random vector**: vector $\boldsymbol{y}$ of random variables
  - **mean (expected value)** $E(\boldsymbol{y}) = (E(y_1), ..., E(y_n))' = \boldsymbol{\mu}$ (*non-random* vector)
    - $E(y_i) = \mu_i$
    - **for a random matrix**: $E(Y) = \begin{pmatrix} E(y_{11}) & \dots & E(y_{1n}) \\ \vdots & \ddots & \vdots \\ E(y_{n1}) & \dots & E(y_{nn}) \end{pmatrix}$ (*non-random* matrix)
  - **properties**: a scalar constant, $\boldsymbol{b}$ vector of constants, $\boldsymbol{y}$ random vector, A matrix of constants...
    - $E(a\boldsymbol{y} + \boldsymbol{b}) = aE(\boldsymbol{y}) + \boldsymbol{b}$
    - $E(A\boldsymbol{y}) = A\ E(\boldsymbol{y})$
    - $E(\boldsymbol{y}'A) = E(\boldsymbol{y})'A$
    - $Var(A\boldsymbol{y}) = A\ Var(\boldsymbol{y})A'$
    - if $\boldsymbol{y}$ is normal distributed, so is $A\boldsymbol{y}$
- **covariance matrix** Σ: diagonal elements $Var(y_i)$, off-diagonal elements $Cov(y_i, y_j)$
  - **formally**: $Var(Y) = \Sigma = \begin{pmatrix} E((y_1 - \mu_1)(y_1 - \mu_1)) & \dots & E((y_1 - \mu_1)(y_n - \mu_n)) \\ \vdots & \ddots & \vdots \\ E((y_n - \mu_n)(y_1 - \mu_1)) & \dots & E((y_n - \mu_n)(y_n - \mu_n)) \end{pmatrix} = E((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})')$
  - Σ symmetric, because $Cov(y_i, y_j) = Cov(y_j, y_i)$
  - Σ diagonal **if** observations (y_i) are independent, because $Cov(y_i, y_j) = 0, i \neq j$

Multiple Regression

- **general linear model**: $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$
  - **response variable** y
  - **several independent (predictor, explanatory) variables** x_i
  - n **cases,** p **predictor values**: $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i = \mu_i + \epsilon_i$
    - x_{ij}: value of the j-th predictor variable of the i-th case
    - $y_1, ..., y_n$ iid., normal distributed, $y_i \sim N(\mu_i, \sigma^2)$

    - μ_i non-random (deterministic)
    - $E(\epsilon_i) = 0$
        - $E(y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$
    - $Var(\epsilon_i) = \sigma^2$
        - $Var(y_i) = \sigma^2$
- **vector form**: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
    - $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $\boldsymbol{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I)$, $E(\boldsymbol{y}) = X\boldsymbol{\beta}$, $Var(\boldsymbol{y}) = \sigma^2 I$
    - $X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}$ fixed, non-random, full rank
    - $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}$
    - $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $Var(\boldsymbol{\epsilon}) = \sigma^2 I$
- **LSE**: $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y}$
    - **fitted values** $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}} = H\boldsymbol{y}$
        - $H = X(X'X)^{-1}X' \in n \times n$
            - H is the orthogonal projection of $\boldsymbol{y}$ onto the linear space spanned by column vectors of X
            - H symmetric ($H' = H$)
            - H idempotent ($HH = H$)
        - $E(\hat{\boldsymbol{y}}) = X\boldsymbol{\beta}$
        - $Var(\hat{\boldsymbol{y}}) = \sigma^2 H$
    - **residuals** $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (I - H)\boldsymbol{y}$
        - $(I - H)$ projects $\boldsymbol{y}$ onto the *perpendicular* space to the linear space spanned by the column vectors of X
        - $(I - H)$ symmetric ($(I - H)' = (I - H)$)
        - $(I - H)$ idempotent ($(I - H)(I - H) = (I - H)$)
        - **rearranged**: $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e} = H\boldsymbol{y} + (I - H)\boldsymbol{y}$
        - $E(\boldsymbol{e}) = \mathbf{0}$
        - $Var(\boldsymbol{e}) = \sigma^2(I - H)$

    - $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
        - $E(\hat{\beta}_i) = \beta_i$
    - $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(X'X)^{-1}$
        - $Var(\hat{\beta}_i) = \sigma^2 v_{ii}$, where v_{ii} is the corresponding diag. el. in $(X'X)^{-1}$
- **MLE**: $s^2 = \frac{SSE}{n-k-1} = \frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (for k predictors *not* including the intercept!)

Source	d.f.	SS (Sum of Squares)	MS (Mean Square)	F
Regression	k	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$
Residual (Error)	$n-k-1$	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-k-1} = s^2$	
Total	$n-1$	$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$		

- **alternative ANOVA calculations**:
    - $SST = \boldsymbol{y'y} - n\overline{\boldsymbol{y}}^2$
    - $SSE = \boldsymbol{y'y} - \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}}$
    - $SSR = SST - SSE = \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}} - n\overline{\boldsymbol{y}}^2$
- **multiple** R^2: "usefulness" of regression...
    - $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ (variation due to regression over total variation)
    - adding a variable to a model increases the regression sum of squares, and hence R^2
        - if adding a variable only marginally increases R^2, it might cast doubt on its inclusion in the model
- **F-test**: $H_0 : \beta_1 = ... = \beta_k = 0$ vs. H_A : at least one $\beta_j \neq 0$
    - **alternative**: $H_{restrict} : E(y) = \beta_0$ vs. $H_{full} : E(y) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
    - $F = \frac{MSR}{MSE} \sim F_{k,n-k-1}$ (bottom of R output)
        - **quantile approach**: reject H_0 if $F > F_{k,n-k-1,1-\alpha}$
        - **probability approach**: reject H_0 if $P(F_{random} > F) < \alpha$ where $F_{random} \sim F_{k,n-k-1}$
- **t-test**: $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$
    - $t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$
        - reject H_0 if $2 \cdot P(T > |t|) < \alpha$ where $T \sim t_{n-k-1}$
        - $100(1-\alpha)\%$ **confidence interval for** β_j: $\hat{\beta}_j \pm t_{n-k-1,1-\alpha/2} \cdot se(\hat{\beta}_j)$
- **linear combination of coefficients**: for when we want to estimate a result *with* given predictors
    - **example (book)**: estimating avg. formaldehyde concentration in homes with UFFI ($x_1 = 1$) and airtightness 5 ($x_2 = 2$)

    - $\theta = \beta_0 + \beta_1 + 5\beta_2 = \boldsymbol{a'\beta}$ with $\boldsymbol{a'} = (1, 1, 5)$
    - **estimate**: $\hat{\theta} = \boldsymbol{a'}\hat{\boldsymbol{\beta}} = (1, 1, 5) \begin{pmatrix} 31.37 \\ 9.31 \\ 2.85 \end{pmatrix} = 54.96$
- **additional sum of squares principle (linear hypthoseses)**: testing *simultaneous* statements about *several* parameters
    - **example**:
        - **full model**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
        - **restrictions**: each restriction is *one* equation equaling 0
            - $\beta_1 = 2\beta_2$ (or $\beta_1 - 2\beta_2 = 0$)
            - $\beta_3 = 0$
        - **matrix form**: matrix $A \in a \times (k+1)$ has one row for each restriction and one column per parameter (+ full rank)
            - $\begin{pmatrix} 0 & 1 & -2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$
            - **hypothesis**: $H_0 : A\boldsymbol{\beta} = \boldsymbol{0}$ vs. H_A : at least one of these $\beta_j \neq 0$
                - **alternative**: $H_{restrict} : \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ vs. $H_{full} : \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$
        - **restricted model**: $y = \beta_0 + \beta_2(2x_1 + x_2) + \epsilon$
    - **additional sum of squares**: $SSE_{restrict} - SSE_{full}$
        - **(!) for** $\mu = \beta_0$: $SSE_{restrict} = SST$
    - **test statistic**: $F = \frac{(SSE_{restrict} - SSE_{full})/a}{SSE_{full}/(n-k-1)} \sim F_{a,n-k-1}$ for a rows in A, k parameters, n observations
        - reject H_0 if p-value $< \alpha$

Specification

- **one-sample problem**: $y_i = \beta_0 + \epsilon_i$
    - $y_1, ..., y_n$ observations taken under uniform conditions from a stable model with mean level β_0
    - $E(y_i) = \beta_0$
    - $E(\boldsymbol{y}) = X\boldsymbol{\beta}$
        - $\boldsymbol{y} = (y_1, ...y_n)'$
        - $X = (1, ..., 1)'$
        - $\boldsymbol{\beta} = \beta_0$

- $\hat{\beta}_0 = \overline{y}$
- $\hat{\sigma}^2 = s^2 = \frac{s_{yy}}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$
- $SSE = SST$
- **in R**: `lm(y~1)`

- **two-sample problem**: $y_i = \begin{cases} \beta_1 + \epsilon_i & i = 1, 2, .., m \\ \beta_2 + \epsilon_1 & i = m+1, ..., n \end{cases}$
  - $y_1, ...y_m$ taken under one set of conditions (standard process), mean β_1
  - $y_{m+1}, ..., y_n$ taken under another set of conditions (new process), mean β_2
  - **alternative**: $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$
    - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2}$
    - x_{i1}, x_{i2} indicator variables
      - $x_{i1} = \begin{cases} 1 & i = 1, 2, ..., m \\ 0 & i = m+1, ..., n \end{cases}$
      - $x_{i2} = \begin{cases} 0 & i = 1, 2, ..., m \\ 1 & i = m+1, ..., n \end{cases}$
    - $E\begin{pmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \beta_1 + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \beta_2$
      - **matrix form** $E(\boldsymbol{y}) = X\boldsymbol{\beta}$
        - $X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}$
        - $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$
  - **in R**: `lm(y~x1+x2-1)`
  - **hypothesis**: $\beta_1 = \beta_2$
- **polynomial models**:
  - **linear**: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

    - $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$
    - `lm(y~x)`
  - **quadratic**: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$
    - $X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$
    - `lm(y~x+I(x^2))`
  - **k-th degree**: $y_i = \beta_0 + \beta_1 x_i + ... + \beta_k x_i^k + \epsilon_i$
    - $X = \begin{pmatrix} 1 & ... & x_1^k \\ \vdots & \ddots & \vdots \\ 1 & ... & x_n^k \end{pmatrix}$
    - `lm(y~poly(x, degree=k, raw=T))`
- **systems of straight lines**: yields of a chemical process which changes linearly with temperature...
  - $y_1, ..., y_m$: yields of a chemical process at temperatures $t_1, ..., t_m$ in the **absence** of a catalyst ($x_i = 0$)
  - $y_{m+1}, ..., y_{2m}$: yields of a chemical process at the same temperatures $t_1, ..., t_m$ in the **presence** of a catalyst ($x_i = 1$)
  - **case a (main effects)**: the catalyst has an effect; the effect is the **same** at all temperatures
    - $\mu_i = \begin{cases} \beta_0 + \beta_1 t_i & i = 1, 2, ..., m \\ \beta_0 + \beta_1 t_{i-m} + \beta_2 & i = m+1, ..., 2m \end{cases}$
    - **alternative (indicator variable)**: $E(y_i) = \beta_0 + \beta_1 t_i + \beta_2 x_i$
      - $x_i = \begin{cases} 0 & i = 1, 2, ..., m \\ 1 & i = m+1, ..., 2m \end{cases}$
      - $t_{i+m} = t_i, i = 1, 2, ..., m$
    - **matrix form**: $E(\boldsymbol{y}) = X\boldsymbol{\beta}$
      - $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_{2m} \end{pmatrix}$

        - $X = \begin{pmatrix} 1 & t_1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & t_m & 0 \\ 1 & t_1 & 1 \\ \vdots & \vdots & \vdots \\ 1 & t_m & 1 \end{pmatrix}$
        - $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$
    - **hypothesis**: $\beta_2 = 0$
- **case b (interaction)**: the catalyst has an effect; the effect **changes** with temperature
    - $\mu_i = \beta_0 + \beta_1 t_i + \beta_2 x_i + \beta_3 t_i x_i \quad i = 1, 2, ..., 2m$
    - **catalyst absent ($x_i = 0$)**: $\mu_i = \beta_0 + \beta_1 t_i \quad i = 1, 2, ..., m$
    - **catalyst present ($x_i = 1$)**: $\mu_i = \beta_0 + \beta_1 t_{i-m} + \beta_2 + \beta_3 t_{i-m} \quad i = m+1, ..., 2m$
        - $\mu_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) t_{i-m}$
    - **matrix form**: $E(\boldsymbol{y}) = X\boldsymbol{\beta}$
        - $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \\ y_{m+1} \\ \vdots \\ y_{2m} \end{pmatrix}$
        - $X = \begin{pmatrix} 1 & t_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & 0 & 0 \\ 1 & t_1 & 1 & t_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_m & 1 & t_m \end{pmatrix}$
        - $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$
    - **hypothesis**: $\beta_2 = \beta_3 = 0$ (no rejection $\rightarrow$ catalyst has no effect)
        - **catalyst depends on temperature?** $\beta_3 = 0$
- **one-way classification (k-sample problem)**: comparison of several "treatments"; generalization of the two-sample problem
    - k catalysts, n_i observations with the i-th catalyst ($i = 1, ..., k$)
        - $n = n_1 + ... + n_k$ total observations

- y_{ij}: j-th observation from the i-th catalyst group ($i = 1, ..., k; \quad j = 1, ..., n_i$)
  - $E(y_{ij}) = \beta_i$
- **matrix form**: $E(\boldsymbol{y}) = X\boldsymbol{\beta} = \beta_1 \boldsymbol{x}_1 + ... + \beta_k \boldsymbol{x}_k$
  - $\boldsymbol{x}_i$: regressor vectors indicating the group membership of the observations
    - $x_{ji} = \begin{cases} 1 & y_{ij} \text{ from group } i \\ 0 & \text{otherwise} \end{cases}$
  - **example (3 groups)**:
    - $\boldsymbol{y} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix}$
    - $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix}$
    - $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$
- **LSE**: $\hat{\beta}_i = \overline{y}_i$
- **hypothesis**: $\beta_1 = \beta_2 = ... = \beta_k$
- **alternative (reference group)**: relate group means to the mean of a reference group (here, the first group)
  - $\beta_i = \beta_1 + \delta_i, \quad i = 2, 3, ...k$
  - $E(y_{ij}) = \begin{cases} \beta_1 & i = 1 \\ \beta_1 + \delta_i & i = 2, ..., k \end{cases}$
  - **matrix form**: $E(\boldsymbol{y}) = X\boldsymbol{\beta}$ where $X = (\boldsymbol{1}, \boldsymbol{x}_2, ..., \boldsymbol{x}_k)$ and $\boldsymbol{\beta} = (\beta_1, \delta_2, ..., \delta_k)'$
    - **example (3 groups)**:

    - $X = (\mathbf{1}, \boldsymbol{x}_2, \boldsymbol{x}_3) = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}$
    - $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \delta_2 \\ \delta_2 \end{pmatrix}$
  - **LSE**: $\hat{\boldsymbol{\beta}} = (\overline{y}_1, \ \overline{y}_2 - \overline{y}_1, \ \ ..., \ \ \overline{y}_k - \overline{y}_1)'$
- **multicollinearity**: in the presence of one variable, the other is not important enough to have it included; the two variables express the same information, so there is no point to include both (p. 157 / 171)
  - typically shown by the fact that, in a model which includes both covariates, *neither* is significant on its own (t-test)
- **orthogonality**: special properties for X matrices with orthogonal columns (dot product of any two columns 0)...
  - **non-changing estimates**: β_i remains the same, regardless of how many variables there are in the model
  - **additivity of SSRs**: $SSR(x_1, ..., x_k) = SSR(x_1) + ... + SSR(x_k)$, for a differing number of variables in a model
  - **orthogonal** $\implies$ **independence**: the components of $\hat{\boldsymbol{\beta}}$ are independent (covariances between β_i zero)

Model Diagnostics

- **possible reasons for a model being inadequate**:
  - **inadequate functional form**: missing needed variables and nonlinear components
  - **incorrect error specification**: non-constant $Var(\epsilon_i)$, non-normal distribution, non-independent errors
  - **unusual observations**: outliers playing a big part
- **residual analysis**: using the residual to assess the adequacy of a model
  - **residual**: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$
    - $\hat{\boldsymbol{y}} = H\boldsymbol{y}$
    - i**-th case in dataset**: $e_i = y_i - \hat{y}_i$

    - estimates the random component $\boldsymbol{\epsilon}$
    - $E(\boldsymbol{e}) = (I - H)E(\boldsymbol{y})$
        - **correctly specified model**: $E(\boldsymbol{e}) = \boldsymbol{0}$
            - $E(\boldsymbol{e}) = (I - H)E(\boldsymbol{y}) = (I - H)X\boldsymbol{\beta} = ... = X\boldsymbol{\beta} - X\boldsymbol{\beta} = \boldsymbol{0}$
        - **incorrectly specified model**: $E(\boldsymbol{e}) \neq \boldsymbol{0}$
            - **"true" model**: $E(\boldsymbol{y}) = X\boldsymbol{\beta} + \boldsymbol{u}\gamma$
                - $\boldsymbol{u}$: regressor vector *not* in $L(X)$
                - γ: a parameter
            - $E(\boldsymbol{e}) = (I - H)E(\boldsymbol{y}) = (I - H)(X\boldsymbol{\beta} + \boldsymbol{u}\gamma) = \gamma(I - H)\boldsymbol{u} \neq 0$
- $\boldsymbol{e}$ and $\hat{\boldsymbol{y}}$ should be **uncorrelated**
    - fitted values should not carry any information on the residuals
    - **in other words**: a graph of the residuals against the fitted values should show *no patterns*
- **properties**: for $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where h_{ij} are elements of H...
    - $Var(\epsilon_i) = \sigma^2$ constant
        - $Var(e_i) = \sigma^2(1 - h_{ii})$ *not* constant
    - $Cov(\epsilon_i, \epsilon_j) = 0, \quad i \neq j$ uncorrelated
        - $Cov(e_i, e_j) = -\sigma^2 h_{ij}, \quad i \neq j$ *not* uncorrelated
- **standardized residuals**: residuals standardized to have approx. mean zero and variance one
    - **definition**: $e_i^s = \frac{e_i}{s}$
        - **recall**: $\hat{\sigma}^2 = s^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n-k-1}$
- [studentized residuals]: the dimensionless ratio resulting from the division of a residual by an estimate of its standard deviation
    - $|d_i| > 2$ or 3 would make us question whether the model is adequate for that case i
    - a histogram or a dot plot of the studentized residuals helps us assess whether one or more of the residuals are unusually large

- [serial correlation (autocorrelation)]: if a regression model is fit to *time series* data (e.g. monthly, yearly...), it is likely that errors are serially correlated (as opposed to the errors ϵ_t being independent for time indices t)
    - **positively autocorrelated**: a positive error last time unit implies a similar positive error this time unit
    - **detection**: calculate *lag k sample autocorrelation* r_k of the residuals ($r_0 = 1$)
        - measures the association within the same series (residuals) k steps apart
        - sample correlation between e_t and its k-th lag, e_{t-k}
        - lag k autocorrelation always between -1 and $+1$

        - **graphically**: plot e_t against e_{t-k} and look for associations (positive: upwards, negative: downwards)
        - **in R**: `acf(fit$residuals,las=1)`
    - **autocorrelation function (of the residuals)**: graph of autocorrelations r_k as a function of the lag k
        - two horizontal bands at $\pm\frac{2}{\sqrt{n}}$ are added to the graph
            - sample autocorrelations that are *outside* these limits are **indications of autocorrelation**
            - if (almost) all autocorrelations are within these limits, one can make the **assumption of independent errors**
    - **Durbin-Watson test**: examines lag 1 autocorrelation r_1 in more detail; complicated to compute
        - $DW \approx 2$: *independent* errors
        - $DW > 2$ or $DW < 2$: *correlated* errors
- **outlier**: an observation that differs from the majority of the cases in the data set
    - one must distinguish among outliers in **the y (response) dimension** (a) vs. outliers in **the x (covariate) dimension** (b) vs. outliers in **both dimensions** (c)
    - x **dimension**: outliers that have unusual values on one or more of the **covariates**
    - y **dimension**: outliers are linked to the **regression model**
        - random component too large?
        - response or covariates recorded incorrectly?
        - missing covariate?
    - **detection**: graphically, studentized residual, *leverage*
- **influence**: an individual case has a **major influence** on a statistical procedure if the **effects** of the analysis are **significantly altered** when the case is **omitted**
- leverage: a measure of how far away the independent variable values of an observation are from those of the other observations
    - **definition**: h_{ii} for i-th independent observation, $i = 1, .., n$ (entry in hat matrix H)
    - **properties**:
        - h_{ii} is a function of the **covariates** (x) but not the response
        - h_{ii} is higher for x farther away from the centroid $\overline{x}$
        - $\sum_{i=1}^{n} h_{ii} = tr(H) = k + 1$
        - $\overline{h} = \frac{k+1}{n}$
        - **rule of thumb**: a case for which **the leverage exceeds twice the average** is considered a **high-leverage case**
            - **formally**: $h_{ii} > 2\overline{h} = \frac{2(k+1)}{n}$
- **influence**: study how the deletion of a case affects the parameter estimates

- **after deleting the i-th case**: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for the remaining $n-1$ cases
- $\hat{\boldsymbol{\beta}}_{(i)}$: the estimate of $\boldsymbol{\beta}$ without the i-th case
- $\hat{\boldsymbol{\beta}}$: the estimate of $\boldsymbol{\beta}$ for all cases
- **influence of the i-th case**: $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$
- Cook's D statistic: estimate of the influence of a data point when performing a least-squares regression analysis
  - $D_i > 0.5$ should be examined
  - $D_i > 1$ great concern

Lack of Fit

- **lack of fit test**: can be performed if there are repeated observations at some of the constellations of the explanatory variables
  - **formally**: for n observations, only k different values of x were observed
    - there were n_i values of y measured at covariate value $x_i, \quad i = 1, ..., k$
    - $x_1 : y_{11}, ..., y_{1n_1}$
      $\vdots$
      $x_k : y_{k1}, ..., y_{kn_k}$
  - **one-way classification model**: $y_{ij} = \beta_1 I(x_i = x_1) + ... + \beta_k I(x_i = x_k) + \epsilon_{ij}$
    - so, in essence, each individual result for a certain constellation x_i is just $\beta_i + \epsilon_{ij}$
    - $E(\epsilon_{ij}) = 0$
    - $Var(\epsilon_{ij}) = \sigma^2$
  - **matrix form**: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
    - $\boldsymbol{y} \in n \times 1$: vector of responses, $n = \sum_{i=1}^{k} n_i$
    - $X \in n \times k$: design matrix with ones and zeros representing the k groups
    - $\boldsymbol{\beta} \in k \times 1$: vector of unknown means μ_i
  - $(\hat{\beta}_1, ..., \hat{\beta}_k) = (\overline{y}_1, ..., \overline{y}_k)$
    - $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ (avg. of group i)
  - **restricted (parametric) model**: $y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$ estimate via least squares
  - **PESS**: $PESS = SSE_{full} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2$
    - **d.f.** number of observations minus number of groups $(n-k)$
  - **LFSS**: $LFSS = \sum_{i=1}^{k} n_i (\overline{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \geq 0$
    - **d.f.** number of groups k - number of parameters (lin: 2 params.)
    - $SSE_{restrict} = PESS + LFSS$

    - $SSE_{restrict} \geq SSE_{full}$
  - **test (linear)**: $H_{restrict}$: $\mu_i = \beta_0 + \beta_1 x_i$ vs. $H_{full} = \mu_i = \beta_1 I(x_i = x_1) + ... + \beta_k I(x_i = x_k)$
    - $F = \frac{LFSS/(k-2)}{PESS/(n-k)} \sim F_{k-2,n-k}$
    - reject $H_0 \implies$ lack of fit, reject restricted model
  - **test (general)**: $H_{restrict}$: $\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ...$ vs. $H_{full} = \mu_i = \beta_1 I(x_i = x_1) + ... + \beta_k I(x_i = x_k)$
    - $F = \frac{LFSS/(k-\dim(\beta))}{PESS/(n-k)} \sim F_{k-\dim(\beta),n-k}$
- variance-stabilizing transformations: find a simple function g to apply to values x in a data set to create new values $y = g(x)$ such that the variability of the values y is not related to their mean value
  - **assume** $y = \mu + \epsilon$ where μ is a **fixed** mean
    - $Var(y) = (h(\mu))^2 \sigma^2$ has a **non-constant variance that depends on the mean**
    - h known
  - **goal**: find $g(x)$ such that $Var(g(x))$ is **constant** and **does not depend on** μ
  - $Var(g(y)) = (g'(\mu))^2 (h(\mu))^2 \sigma^2$
    - **goal**: find g **such that** $g'(\mu) = \frac{1}{h(\mu)}$, such that finally $Var(g(y)) \approx \sigma^2$
    - **example**: $h(\mu) = \mu \implies g'(\mu) = \frac{1}{\mu} \implies g(\mu) = \ln(\mu)$
  - **Box-Cox transformations**: find λ s.t. the transformed response $y_i^{(\lambda)}$ **minimizes** $SSE(\lambda)$ (done in a table, compare various λs to their $SSE(\lambda)$s)
    - $\lambda = 0$: log transform using limit `lm(log(...)~., data=...)`
    - $\lambda = \frac{1}{2}$: square root transform `lm(sqrt(...)~., data=...)`
    - $\lambda = 1$: no transform `lm(...~., data=...)`
    - $\lambda = 2$: square transform `lm((...)*(...)~., data=...)`
    - **in R**: `library(MASS); boxcox(fit)`

Model Selection

- **goal**: given *observational data*, find the **best** model which incorporates the concepts of **model fit** and **model simplicity**
  - increasing the number of **predictors** increases **variability** in the predictions
    - $Var(\hat{\boldsymbol{y}}) = \sigma^2 H \implies$ average variance $\frac{1}{n} \sum_{i=1}^{n} Var(\hat{\boldsymbol{y}}) = \frac{\sigma^2 (k+1)}{n}$ for k covariates and sample size n
  - **multicollinearity**: different methods of analysis may end up with final models that **look very different, but describe the data equally well**

- **model selection**: given observations on a repsonse y and q potential explanatory variables $v_1, ..., v_q$, select a model $y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$ where...
  - $x_1, ..., x_p$ is a subset of the original regressors $v_1, ..., v_q$
  - no important variable is left out of the model
  - no unimportant variable is included in the model
- **all possible regressions**: fit 2^q models if q variables are involved
  - $R_k^2 = 1 - \frac{SSE_k}{SST}$ for k variables and $k + 1$ regression coefficients
    - **increase** in k means **decrease** in SSE_k, approaching 0 when $k = n - 1$
      - **increase** in R^2 means **decrease** in s^2
    - SST does not depend on the covariates; just on y (see definition)
    - therefore, R^2 approaches 1 as k increases, so we don't use R^2, since we'd just choose the model with the most variables
  - R_{adj}^2: adjusted R^2
    - remedies the problem of R^2 continually increasing by dividing the degrees of freedom
    - **ideal model and choice of k**: highest R_{adj}^2
      - **equivalent**: smallest s^2
  - AIC: Akaike's Information Criterion
    - prefer models with **smaller** AIC
  - BIC: Bayesian Information Criterion
    - larger penalty for more variables
- **automatic model selection methods**: forward selection, backward elimination, stepwise regression (needed in R: `library(MASS)`)
  - **forward selection**: start with the smallest model, build up to the optimal model
    - **in R**: `stepAIC(lm(y~1, data=dataset), direction = "forward", scope = list(upper = lm(y~., data=dataset))[, k = log(nrow(dataset))])` (for BIC: `[...]`)
  - **backward elimination**: start with the largest model and build down to the optimal model
    - **in R**: `stepAIC(lm(y~., data=dataset), direction = "backward"[, k = log(nrow(dataset))])`
  - **stepwise regression**: oscillate between forward selection and backward elimination
    - **in R**: use either previous function, but with `direction = "both"`

```
forward selection
```

```
INIT
  M = intercept-only model
  P = all covariates

REPEAT
  IF P empty STOP
  ELSE
    calculate AIC for sizeof(P) models, each model containing one covariate
in P is added to M
    IF all AICs > AIC(M) STOP
    ELSE
      update M with covariate whose addition had minimum AIC
      remove covariate from P
```

```
backward elimination

INIT
  M = model with all covariates
  P = all covariates

REPEAT
  IF P empty STOP
  ELSE
    calculate AIC for sizeof(P) models, each model without each of the
covariates in P
    IF all AICs > AIC(M) STOP
    ELSE
      update M by deleting covariate that led to minimum AIC
      remove covariate from P
```

```
stepwise regression

INIT
  M = intercept-only model OR full model
  e = small threshold

REPEAT UNTIL STOP
  do a forward step on M
  do a backward step on M

# For both steps, the differences in AIC need to be
# greater than e for the selection to go forward, otherwise
# the changes can keep undoing each other.
```

Nonlinear Regression

- **linear model (recap)**: $y = \mu + \epsilon$ where $\mu = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
  - **key**: linearity of the parameters β_i
  - the regressor variables x_i can be any known nonlinear function of the regressors...
- **intrinsically nonlinear model**: a nonlinear model that cannot be transformed into a linear model
  - **counterexample**: $y = \alpha x_1^\beta x_2^\gamma \epsilon$ can be transformed into $\ln(y) = \ln(\alpha) + \beta \ln(x_1) + \gamma \ln(x_2) + \ln(\epsilon)$
  - require *iterative* algorithms and convergence (vs. linear models, which are *analytic*)
- **linear trend model**: $\mu_t = \alpha + \gamma t$
  - γ: growth rate, unbounded
  - α: starting value at $t = 0$
- **nonlinear regression model**: $y_i = \mu_i + \epsilon_i = \mu(\boldsymbol{x}_i, \boldsymbol{\beta}) + \epsilon_i$
  - $\epsilon_i \sim N(0, \sigma^2)$ iid. for $i = 1, ..., n$
  - $\boldsymbol{x}_i = (x_{i1}, ..., x_{im})'$: vector of m covariates for the i-th case (typically $m = 1$, where the covariate is time)
  - $\boldsymbol{\beta}$: vector of p parameters to be estimated along with σ^2 (usually a different num. of parameters than covariates)
  - $\mu(\boldsymbol{x}_i, \boldsymbol{\beta})$: nonlinear model component
  - $S(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n (y_i - \mu(\boldsymbol{x}_i, \boldsymbol{\beta}))^2$
  - **estimates**:
    - $\hat{\boldsymbol{\beta}}$: no closed form!
      - use iterative function to minimize $S(\hat{\boldsymbol{\beta}})$
    - $\hat{\sigma}^2 = s^2 = \frac{S(\hat{\boldsymbol{\beta}})}{n-p} = \frac{\sum_{i=1}^n (y_i - \mu(\boldsymbol{x}_i, \boldsymbol{\beta}))^2}{n-p}$
  - $Var(\hat{\boldsymbol{\beta}}) \approx s^2 (X'X)^{-1}$
    - $s.e.(\hat{\beta}_i) = \sqrt{v_{ii}}$ (the square roots of the diagonal elements in the covariance matrix provide estimates of the standard errors)
    - off-diagonal elements provide estimates of the covariances among the estimates
  - $100(1-\alpha)\%$ **C.I. for** β_j: $\hat{\beta}_j \pm t_{n-p,1-\alpha/2} s.e.(\hat{\beta}_j)$
  - $H_0 : \beta_j = 0$ vs. $H_a : \beta_j \neq 0$: $t = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)} \sim t_{n-p}$
    - **for constants**: $H_0 : \beta_j = c$ vs. $H_a : \beta_j \neq c$: $t = \frac{\hat{\beta}_j - c}{s.e.(\hat{\beta}_j)} \sim t_{n-p}$
  - restricted models or goodness-of-fit tests can be performed similarly as for linear models
  - **in R**: `fitnls = nls(y~(formula using a, b),start=list(a=...,b=...))`

- **Newton-Raphson method**: a root-finding algorithm which produces successively better approximations to the zeroes of a real-valued function
  - **goal**: find $\boldsymbol{\beta}$ that minimizes $f(\boldsymbol{\beta})$, here $S(\boldsymbol{\beta})$ or $-\log L(\boldsymbol{\beta})$
    - $Df(\boldsymbol{\beta}) = (\frac{\partial f}{\partial \beta_1}, ..., \frac{\partial f}{\partial \beta_p})'$: the p-vector containing the first derivatives of f w.r.t. β_i
    - $D^2 f(\boldsymbol{\beta})$: the $p \times p$ matrix of second derivatives with the ij-th element $\frac{\partial^2 f}{\partial \beta_i \partial \beta_f}$ (Hessian matrix)
  - **general**: initialize $\boldsymbol{\beta}_{old} =$ starting value, then repeat until convergence:
    1. $\boldsymbol{\beta}_{new} \approx \boldsymbol{\beta}_{old} - (D^2 f(\boldsymbol{\beta}_{old}))^{-1} Df(\boldsymbol{\beta}_{old})$
    2. $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}$
    - **problem**: unstable due to inversion
  - **scoring**: initialize $\boldsymbol{\beta}_{old} =$ starting value, then repeat until convergence:
    1. $\boldsymbol{\beta}_{new} \approx \boldsymbol{\beta}_{old} + (I(\boldsymbol{\beta}_{old}))^{-1} D \log L(\boldsymbol{\beta}_{old})$
    2. $\boldsymbol{\beta}_{old} = \boldsymbol{\beta}_{new}$
    - **information matrix**: $I(\boldsymbol{\beta}) = E(-D^2 \log L(\boldsymbol{\beta}))$
  - **problematic**: local minima that "trap" iterative algorithms, parameters of highly varying magnitudes (e.g. one parameter in range 0-1, another in the thousands), badly specified models with non-identifiable parameters (similar to multicollinearity)

Time Series Models

- **first-order autoregressive model (AR1)**: $y_t = \mu(X_t, \beta) + \epsilon_t$
  - **autocorrelations of observations 1 step apart**: ϕ
    - all correlations among observations one step apart ϕ are the same
    - $\phi = Corr(\epsilon_1, \epsilon_2) = ... = Corr(\epsilon_{n-1}, \epsilon_n)$
      - $|\phi| < 1$ (correlations between -1 and $+1$)
  - **autocorrelations of observations k steps apart**: ϕ^k
    - $\phi^k = Corr(\epsilon_1, \epsilon_{k+1}) = ... = Corr(\epsilon_{n-k}, \epsilon_n)$
  - **properties of autocorrelations**:
    - they depend only on the time lag between the observations (so the time indices don't matter; just the time distance)
    - they decrease exponentially with the time lag (because $-1 < \phi < 1$)
      - the farther apart the observations, the weaker the autocorrelation
      - if ϕ is close to 1, the decay is slow
  - **autocorrelation functions (lag k correlation)**: $\rho_k = \phi^k = Corr(\epsilon_{t-k}, \epsilon_t)$
    - $\rho_0 = 1$

        - $\rho_k = \rho_{-k}$
    - **correlated error at time** t: $\epsilon_t = \phi\epsilon_{t-1} + a_t$ where $a_t \sim N(0, \sigma_a^2)$
        - **white noise (random shocks)**: a_t
            - a_t is the "usual" regression model error; mean 0, all uncorrelated
                - $Corr(a_{t-k}, a_t) = 0$ for all $k \neq 0$
        - **expanded**: $\epsilon_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + ...$
        - $E(\epsilon_t) = 0$
        - $Var(\epsilon_t) \to \frac{\sigma_a^2}{1-\phi^2}$
    - **stationary model**: fixed level 0; realizations scatter around the fixed level and sample paths don't leave this level for long periods
    - **in R**: `library(nlme); fitgls=gls(y~x,correlation=corARMA(p=1,q=0))` (change `p=2` for AR2)
- **random walk model**: $y_t = \mu(X_t, \beta) + \epsilon_t$
    - $\phi = 1$
    - $\epsilon_t = \epsilon_{t-1} + a_t = a_t + a_{t-1} + a_{t-2} + ...$
        - cumulative sum of all random shocks up to time t
    - **nonstationary model**: no fixed level; paths can deviate for long periods from the starting point
    - **first-order difference**: $w_t = \epsilon_t - \epsilon_{t-1} = a_t$
        - well-behaved, stationary, uncorrelated
- **effects of ignoring autocorrelation**: what happens when we fit a standard linear model even if the errors are correlated?
    - **stationary errors**: variance of $\hat{\beta}$ will be **overestimated** compared to the true variance (**inefficiency**)
        - t ratios too small $\to$ null hypothesis **less likely** to be rejected when it **should** be
    - **non-stationary errors**: variance of $\hat{\beta}$ will be **underestimated** compared to the true variance
        - t ratios too large $\to$ null hypothesis **likely** to be rejected when it **shouldn't** be
- **forecasting (prediciton)**: given data up to time period n, predict response at time period $n + r$ (r step-ahead forecast)
    - r **step-ahead forecast**: $y_n(r) = \hat{y}_{n+r}$
        - n: forecast origin
        - r: forecast horizon
    - **assumption**: future values of the covariate x_t are known (e.g. own future investments)
    - **1 step forecast (AR1, one covariate)**: assume x_{n+1} known...

    - **observation**: $y_{n+1} = \phi y_n + (1-\phi)\beta_0 + (x_{n+1} - \phi x_n)\beta_1 + a_{n+1}$
    - **prediction**: $\hat{y}_{n+1} = \hat{\phi} y_n + (1-\hat{\phi})\hat{\beta}_0 + (x_{n+1} - \hat{\phi} x_n)\hat{\beta}_1$
        - **95% CI**: $\hat{y}_{n+1} \pm 1.96 se(\hat{y}_{n+1})$
- **general step forecast (AR1, one covariate)**: $\hat{y}_{n+r} = \hat{\phi}\hat{y}_{n+r-1} + (1-\hat{\phi})\hat{\beta}_0 + (x_{n+r} - \hat{\phi} x_{n+r-1})\hat{\beta}_1$ for $r \geq 2$
    - **95% CI**: $\hat{y}_{n+r} \pm 1.96 se(\hat{y}_{n+r})$

Logistic Regression

- logistic regression: regression where the response variable is **binary** (general: **categorical**)
    - y_i: **outcome** of case i, $\quad i = 1, 2, ..., n$
        - $y_i \sim Ber(\pi)$, independent
            - $P(y_i = 1) = \pi$ (success)
                - $\ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = x_i'\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$
                    - $\pi(x_i) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$
                    - $1 - \pi(x_i) = \frac{1}{1+e^{x_i'\beta}}$
                    - β_0: inflection point
                    - β_1: steepness of sigmoid-like function
                - **risk** of y for factor x: $\pi(x) = P(y=1|x) = \frac{e^{x'\beta}}{1+e^{x'\beta}}$
                - **odds** of y for a fixed x: $Odds(x) = \frac{\pi(x)}{1-\pi(x)} = \frac{P(y=1|x)}{1-P(y=1|x)} = \exp(x'\beta)$
                    - how much higher is the probability of the occurrence y compared to the nonoccurrence of y?
                    - odds of $n:1 \implies$ occurence is n times more likely than nonoccurence
                - **odds ratio**: $OR = \frac{Odds(x=1)}{Odds(x=0)} = \frac{\frac{P(y=1|x=1)}{P(y=0|x=1)}}{\frac{P(y=1|x=0)}{P(y=0|x=0)}}$
                    - $\beta = \ln\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \ln(OR)$ vector of log odds ratios
                        - $\exp(\beta) = OR$
                        - what is the multiplicative factor by which the odds of occurrence increase / decrease for a change from x to $x+1$?
                            - e.g. $\beta = -0.2 \to \exp(\beta) = 0.82 \implies$ a change from x to $x+1$ decreases the odds of occurence by 18%
                        - **for k units**: βk with ratio measured as $\exp(\beta k)$
            - $P(y_i = 0) = 1 - \pi$ (failure)
            - $E(y_i) = \pi$

- **one covariate model**: $\exp(x'\beta) = \beta_0 + \beta_1 x$
  - $\ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) = \beta_0 + \beta_1 x$
  - $Odds(x) = \frac{P(y=1|x)}{1-P(y=1|x)} = \exp(\beta_0 + \beta_1 x)$
    - $Odds(x=0) = \exp(\beta_0 + \beta_1 \cdot 0) = \exp(\beta_0)$
    - $Odds(x=1) = \exp(\beta_0 + \beta_1 \cdot 1) = \exp(\beta_0 + \beta_1)$
  - $OR = \frac{Odds(x=1)}{Odds(x=0)} = \exp(\beta_1)$
  - $\beta_1 = \ln(OR)$
    - $H_0 : \beta_1 = 0$ (no assoc. between x and y)
  - $\beta_0 = \ln(Odds(x=0))$
- **MLE** $\hat{\beta}$: Newton-Raphson...
- **CIs and tests**:
  - $100(1-\alpha)\%$ **CI for** $\ln(OR)$: $\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j)$
  - $100(1-\alpha)\%$ **CI for** OR: $\exp(\hat{\beta}_j \pm z_{1-\alpha/2} se(\hat{\beta}_j))$
  - **Wald test**: $H_0 : \beta_j = 0$ vs. $H_A : \beta_j \neq 0$: $\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim N(0,1)$
- **case**: an individual observation
- **constellation**: grouped information at distinct levels of the explanatory variables
  - n_k; number of cases at the k-th constellation
  - y_k: number of successes at the k-th constellation
  - **prob. of success for** k**-th constellation**: $\pi(x_k, \beta) = \frac{\exp(x_k\beta)}{1+\exp(x_k\beta)}$
- **likelihood ratio tests (LRT)**: used to compare the maximum likelihood under the current model (the “full” model), with the maximum likelihood obtained under alternative competing models ("restricted" models)
  - $H_{restrict}$: linear predictor $x'_{res}\beta_{res}$ vs. H_{full} : linear predictor $x'\beta$
    - x_{res} subset of x
  - **LRT statistic**: $2 \cdot \ln\left(\frac{L(full)}{L(restrict)}\right) = 2 \cdot \ln\left(\frac{L(\hat{\beta})}{L(\hat{\beta}_{res})}\right) \sim \chi^2_a$
    - **equiv.**: $2 \cdot (\ln(L(full)) - \ln(L(restrict)))$
    - $a = \dim(\beta) - \dim(\beta_{res})$
    - reject $H_{restrict}$ if statistic greater than corresponding chi-square value
    - **large value** $\implies$ the success probability depends on one or more of the regressors (i.e. full model better)
    - **small value** $\implies$ none of the regressors in the model influence the success probability
- **deviance**: twice the log-likelihood ratio between the saturated model and the parameterized (full) model; m constellations

    - **saturated model**: each constellation of the explanatory variables is allowed its own distinct success probability
        - $\hat{\pi}_k = \frac{y_k}{n_k}$
    - $D = 2\frac{\ln(L(saturated))}{\ln(L(full))} = 2 \cdot \ln\left(\frac{L(\hat{\pi}_1,...,\hat{\pi}_m)}{L(\hat{\beta})}\right) \sim \chi^2_a$
        - $a = m - \dim(\beta)$
        - $LRT = D(restricted) - D(full)$
    - **p-value**: `1 - pchisq(D, a)`
- **in R**: `freqs <- cbind(yes, no); fit <- glm(freqs~x[+...], family="binomial")`

Poisson Regression

- [generalized linear model (GLM)]: generalizes linear regression by allowing the linear model to be related to the response variable via a link function
    - **response variables** $y_1, ...y_,n$: share the same distribution from the *exponential* family (Normal, Poisson, Binomial...)
    - **parameters** β and **explanatory variables** $x_1, ...x_p$
    - **monotone link function** g: relates a transform of the mean μ_i linearly to the explanatory variables
        - $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$
        - **standard linear regression**: $g(\mu) = \mu$ (identity function)
        - **logistic regression**: $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$ (logit)
        - **Poisson regression**: $g(\mu) = \ln(\mu)$
- **Poisson regression model**: response represents **count data** (e.g. number of daily equipment failures, weekly traffic fatalities...)
    - $P(Y = y) = \frac{\mu^y}{y!}e^{-\mu}, \quad y = 0, 1, 2, ...$
    - $E(y) = Var(y) = \mu > 0$
    - $g(\mu) = \ln(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$
        - $\mu = \exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)$
    - **interpretation of coefficients**: changing x_i by one unit to $x_i + 1$ while keeping all other regressors fixed affects the mean of the response by $100(\exp(\beta_i) - 1)\%$
        - **example**: $\frac{\exp(\beta_0+\beta_1(x_1+1)+...+\beta_p x_p)}{\exp(\beta_0+\beta_1 x_1+...+\beta_p x_p)} = \exp(\beta_1)$
    - **95% CI**: $\hat{\beta} \pm 1.96se(\hat{\beta})$
        - **for the mean ratio** $\exp(\beta)$: $\exp(\hat{\beta} \pm 1.96se(\hat{\beta}))$
    - everything else identical to logistic regression

Linear Mixed Effects Models

- **linear mixed effects models**: some subset of regression parameters vary randomly from one individual to another
  - individuals are assumed to have their own subject-specific mean response trajectories over time
  - **simple mixed effects model**: $Y_{ij} = \beta + b_i + e_{ij}$ (observation = population mean + individual deviation + measurement error)
    - β: **population mean** (*fixed effects*, constant)
    - b_i: **individual deviation** from the population mean (*random effects*) (i-th individual)
      - $b_i \sim N(0, d)$
      - **positive**: individual responds **higher** than population average (higher on y-axis)
      - **negative**: individual responds **lower** than population average (lower on y-axis)
    - e_{ij}: **within-individual deviations** (measurement error) (i-th individual, j-th observation)
      - $e_{ij} \sim N(0, \sigma^2)$
    - $E(Y_{ij}) = \beta$
    - $Var(Y_{ij}) = d + \sigma^2$
    - $Cov(Y_{ij}, Y_{km}) = 0$ for $i \neq k$
    - $Cov(Y_{ij}, Y_{ij}) = Var(Y_{ij}) = d + \sigma^2$
    - $Cov(Y_{ij}, Y_{ik}) = d$
    - $Cor(Y_{ij}, Y_{km}) = 0$ for $i \neq k$
    - $Cor(Y_{ij}, Y_{ij}) = 1$
    - $Cor(Y_{ij}, Y_{ik}) = \frac{d}{d+\sigma^2}$
    - **in R**: matrix format, n rows for n individuals and m columns for m time points per individual
      - has to be transformed into longitudinal format for use with `lme`
  - **general linear mixed effects model**: $Y_i = X_i\beta + Z_i b_i + e_i, \;\; b_i \sim N_q(0, D), \;\; e_i \sim N_{n_i}(0, R_i)$, where $R_i = \sigma^2 I_{n_i}$
    - $Y_i \in \mathbb{R}^{n_i}$: outcomes (for i-th individual, as for nearly everything here...)
    - $X_i \in \mathbb{R}^{n_i \times p}$: design matrix for *fixed effects*
    - $Z_i \in \mathbb{R}^{n_i \times q}$: design matrix for *random effects* (columns are usually subset of columns of X_i)
    - $\beta \in \mathbb{R}^p$: fixed effects
      - any component of β can be allowed to vary randomly by including the corresponding column of X_i in Z_i
    - $b_i \in \mathbb{R}^q$: random effects

        - independent of covariates X_i
    - $e_i \in \mathbb{R}^{n_i}$: within-individual errors
    - **conditional mean**: $E(Y_i \mid b_i) = X_i\beta + Z_ib_i$
    - **marginal mean**: $E(Y_i) = X_i\beta$
    - **conditional variance**: $Var(Y_i \mid b_i) = R_i$
    - **marginal variance**: $Var(Y_i) = Z_iDZ_i' + R_i$
    - $\sigma^2_{REML} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (unbiased, restricted maximum likelihood)
    - $\sigma^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (biased)
    - $\hat{b}_i$: best linear unbiased predictor (BLUP)
        - "shrinks" the i-th individual's predicted response profile towards the population-averaged mean response profile
        - large R_i compared to $D \implies$ more shrinkage to mean
        - small R_i compared to $D \implies$ closer to observed value
        - large $n_i \implies$ less shrinkage

Statistical Learning (Machine Learning)

- **supervised learning**: an outcome (which guides the learning process) predicted based on a set of *features*
    - **unsupervised learning**: no outcome; only features observed (not relevant here)
    - **outcome (outputs, responses, dependent variables)**: the thing to predict; can be *quantitative* (ordered, e.g. stock price) or *qualitative* (unordered here, categorical, factors, e.g species of Iris)
        - predicting quantitative outcomes $\rightarrow$ *regression*
        - predicting qualitative outcomes $\rightarrow$ *classification*
    - **features (inputs, predictors, independent variables)**: the data to make predcitions for the outcome
    - **training set**: data set containing *both features and outcomes* to build the model

Prediction Methods

- **least squares model (linear model)**: *high* stability but *low* accuracy (**high bias, low variance**)
    - **goal**: predict Y by $f(X) = X'\beta$
        - β_0: intercept (bias)
        - $X \in \mathbb{R}^p$: random input vector (first element 1 for intercept)
        - $Y \in \mathbb{R}$: random outcome (to predict)
        - $p(X, Y)$: joint distribution
        - $f(X)$: function for predicting Y based on X

        - $f'(X) = \beta \in \mathbb{R}^p$: vector that points in the steepest uphill direction
      - $(x_1, y_1), ..., (x_n, y_n)$: training data
    - **method**: pick β to minimize residual sum of squares $RSS(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2$
      - **matrix notation**: $RSS(\beta) = (y - X\beta)'(y - X\beta)$
    - **LSE**: $\hat{\beta} = (X'X)^{-1}X'y$
      - **fitted value**: $\hat{y}_i = x_i'\hat{\beta}$
    - **theoretical**: $\beta = (E(X'X))^{-1}E(X'Y)$
    - should only be used when Y is continuous and Normal distributed
- k**-nearest neighbor model**: *low* stability but *high* accuracy (**low bias, high variance**)
    - use observations in training set closest in input space to x to form $\hat{Y}$
      - **formally**: $f(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i$
        - $N_k(x)$: **neighborhood** of x defined by the k closest points x_i in the training sample
        - find k observations with x_i closest to new x in input space, and average their responses

Statistical Decision Theory

- **loss function** $L(Y, f(X))$: penalizes errors in prediction
    - **expected prediction error**: $EPE(f) = E_{x,y}(L(Y, f(X)))$
- **expected (squared) prediction error**: criterion for choosing f based on the squared error loss function (L2)
    - **L2 Loss**: $L_2 = L(Y, f(X)) = (Y - f(X))^2$ (most popular)
    - $EPE(f) = E_{x,y}(Y - f(X))^2 = \iint (y - f(x))^2 p(x, y) dx dy = \int (\int (Y - f(x))^2 p(y|x) dy) p(x) dx = E_x(E_{y|x}((Y - f(X))^2 | X))$
    - **optimal Bayes classifier**: minimize $E_{y|x}((Y - f(X))^2 \mid X)$ for all X
      - $f_{bayes}(X) = E_{y|x}(Y|X)$
- **nearest neighbor**: $f(x) = Ave(y_i \mid x_i \in N_k(x))$
    - as $n, k \to \infty$, $\frac{k}{n} \to 0$: $f(x) \to E(Y \mid X = x)$
- **L1 Loss**: $E|Y - f(X)|$ (abs. value)
    - $f(x) = \text{median}(Y \mid X = x)$

Categorical Data

- **estimate** $\hat{G}$: contains values in the set of possible classes $\mathcal{G}$ where $|\mathcal{G}| = K$
- **loss function**: $L \in \mathbb{R}^{K \times K}$
    - zero on he diagonal
    - nonnegative elsewhere

    - $L(k, l)$: price paid for classifying an observation belonging to class $\mathcal{G}_k$ as $\mathcal{G}_l$
  - **zero-one loss function**: all misclassifications are charged one unit
    - $L(k, l) = \begin{cases} 0 & k = l \\ 1 & k \neq l \end{cases}$
  - $EPE = E(L(G, \hat{G}(X))) = E_x(E_{g|x}(L(G, \hat{G}(X)) \mid X))$
    - $E_{g|x}(L(G, \hat{G}(X)) \mid X) = \sum_{k=1}^{K} L(\mathcal{G}_k, f(X)) p(\mathcal{G}_k \mid X)$
      - $\sum_{k=1}^{K} p(\mathcal{G}_k \mid X) = 1$
    - $f(X) = \arg\min_{g \in G} \sum_{k=1}^{K} L(\mathcal{G}_k, g) p(\mathcal{G}_k \mid X)$
    - **zero-one loss**: $f(X) = \arg\max_{g \in G} p(g \mid X)$
    - $f(X) = f_{bayes}(X)$
      - $EPE(f_{bayes}) = E_{x,y}(L(Y, f_{bayes}(X)))$

Summary by Flavius Schmidt, ge83pux, 2025.
https://home.cit.tum.de/~scfl/
Images from Wikimedia.