Statistical Principals of Data Analysis

Dr. Felipe Teles

Graz - Austria, 2023

Table of Contents

Linear Models

A linear model in statistics is a model of the type

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \beta_p X_p + \varepsilon, \tag{1}$$

where Y is a response variable, $X_1, X_2, \ldots, X_{p-1}, X_p$ are explanatory variables, $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}, \beta_p$ are the coefficients of the linear model and $\varepsilon \sim N(0, \sigma^2)$ is the error of the model.

Linear Models

Or, in matrix notation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & \dots & X_{(p-1)1} & X_{p1} \\ 1 & X_{12} & X_{22} & X_{32} & \dots & X_{(p-1)2} & X_{p2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & X_{3n} & \dots & X_{(p-1)n} & X_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad (2)$$

which can be written as

$$Y = X\beta + \varepsilon, \quad (3)$$

where X is the design matrix (explanatory variables, treatments, groups, etc.).

Linear Models

Examples of linear models:

- $Y = \beta_0 + \beta_1 X_1$;
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^5$;
- $Y = \dfrac{\beta_1}{X_1}$;
- $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 X_3$;

Linear Models

Examples of nonlinear models:

- $Y = \beta_0 + \beta_1 X_1 + \dfrac{\beta_2}{\beta_3 + X_1}$;
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\beta_3 X_3}$;
- $Y = cos(\beta_1 X_1) + sin(\beta_2 X_1)$;
- $Y = \beta_0 e^{\beta_1 X1}$;
- $\dfrac{dS}{dt} = -\dfrac{\mu_{max} S}{K_S + S} X Y_{X/S}$, where μ_{max}, K_S, $Y_{X/S}$ are parameters;

Least Squares Method

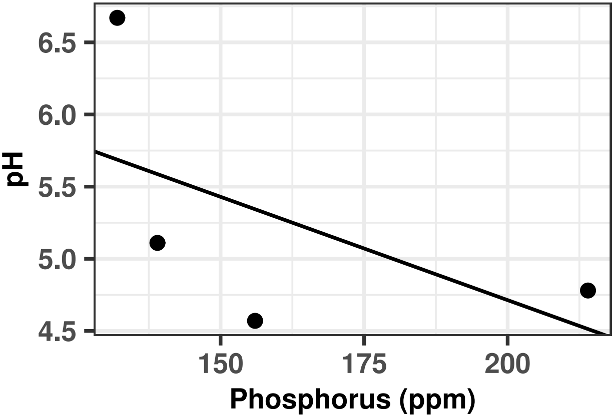

Figure: Linear Model: pH explained by nitrogen from Soil data set

Least Squares Method

Goal: minimize the residual sum of squares (RSS)

$$
\begin{aligned}
RSS &= ||Y - X\beta||^2 && (4)\\
&= (Y - X\beta)^T(Y - X\beta) && (5)\\
&= Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta && (6)\\
&= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta. && (7)
\end{aligned}
$$

Least Squares Method

Goal: minimize the residual sum of squares (RSS)

$$RSS = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$$

$$\min_{\beta} RSS \implies \frac{\partial}{\partial\beta} RSS = 0 \tag{8}$$

$$\begin{aligned}
\frac{\partial}{\partial\beta} RSS &= 0 && (9)\\
-2X^TY + 2X^TX\beta &= 0 && (10)\\
X^TX\beta &= X^TY && (11)\\
(X^TX)^{-1}X^TX\beta &= (X^TX)^{-1}X^TY && (12)\\
\beta &= (X^TX)^{-1}X^TY. && (13)
\end{aligned}$$

$$\hat{\beta} = (X^TX)^{-1}X^TY. \tag{14}$$

Exercise

1. In *R*, install and load the packages *car*, *carData* and *ggplot2*.
2. Load the data set *Soils* and, by using the function *subset*, filter the data for Gp=S1.
3. Fit a linear model for pH using the explanatory variable P.
4. Check the results with *summary*.
5. Theoretical check - calculate $\hat{\beta}$ using Equation (14).

Exercise

```
data01=subset(Soils,Gp=="S1")
model1=lm(data=data01,pH~P)
summary(model1)

X=as.matrix(model.matrix(model1))
Y=as.matrix(data01$pH)
Beta=solve(t(X)%*%X)%*%t(X)%*%Y

ggplot(data=data01,aes(x=P,y=pH))+
geom_point(pch=19)+theme_bw()+
geom_abline(intercept=coef(model1)[1],
slope=coef(model1)[2])
```

Goodness of Fit

Based on what can we say if one model is good?

1. Coefficient of Determination R^2;
2. Distribution of Errors;
3. Unusual observations;
4. Structure of the model.

Goodness of Fit - Coefficient of Determination

The coefficient of determination is the percentage of data variability explained by the model:

$$R^2 = 1 - \frac{\sum_{i=i}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=i}^{n}(y_i - \bar{y})^2} \tag{15}$$

Some

1. $0 \leq R^2 \leq 1$;
2. If $R^2 = 1$, then $RSS = 0$;
3. R^2 is equivalent to the percentage of variability explained regarding total variability.

Goodness of Fit - Distribution of Errors

A linear model:

$$Y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2). \tag{16}$$

Check:

1. Distribution of errors around zero: $E[\varepsilon] = 0$;
2. Homoscedasticity: $Var[\varepsilon] = \sigma^2$;
3. Normality of residuals: Q-Q plot;

Goodness of Fit - Distribution of Errors

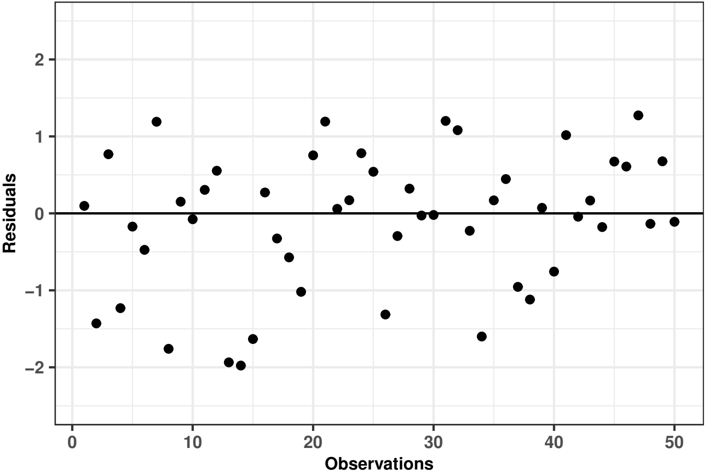

Figure: Residuals normally distributed with homoscedastic variance structure.

Goodness of Fit - Distribution of Errors

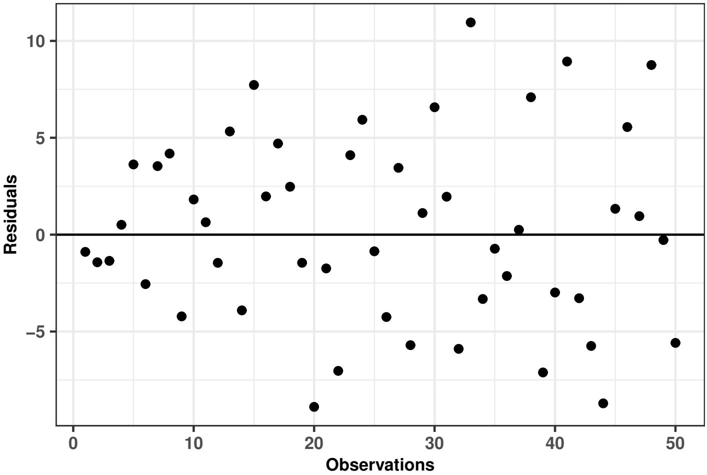

Figure: Residuals normally distributed with mild disturbance.

Goodness of Fit - Distribution of Errors

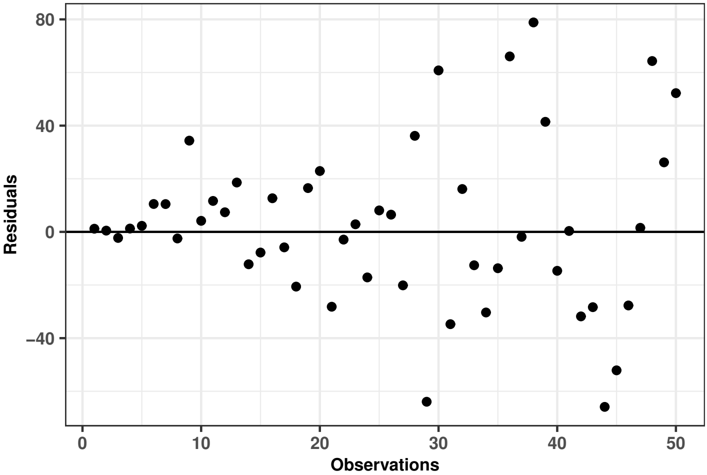

Figure: Residuals normally distributed with heteroscedastic variance structure.

Goodness of Fit - Distribution of Errors

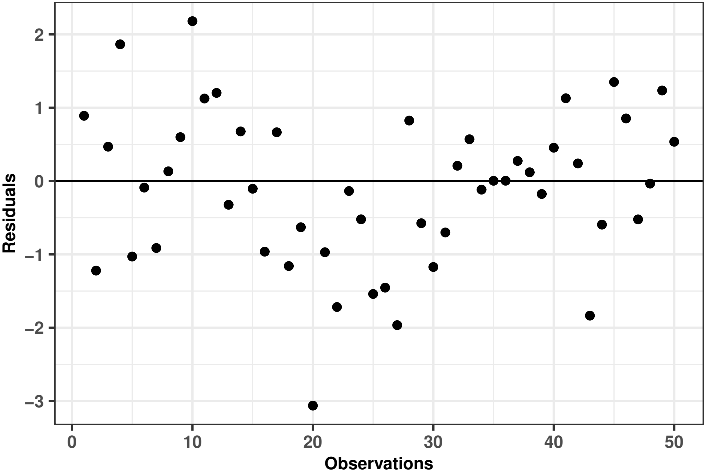

Figure: Nonlinear residuals.

Goodness of Fit - Distribution of Errors

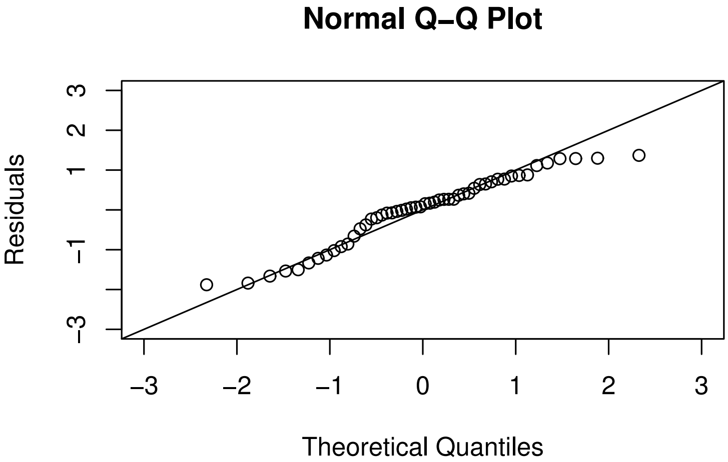

Figure: Q-Q plot.

Goodness of Fit - Distribution of Errors

Test of normality - Shapiro-Wilk test (shapiro.test)

$$H_0 : \varepsilon \sim N(\mu, \sigma^2) \tag{17}$$

$$H_1 : \varepsilon \nsim N(\mu, \sigma^2) \tag{18}$$

Exercise

Using the data set *Soils*:

1. Fit the following model:
   $pH = \beta_0 + \beta_1 N + \beta_2 P + \beta_3 Ca + \beta_4 Mg + \beta_5 K + \beta_6 Na + \varepsilon.$
2. Make the analysis of diagnostics - plot the residuals and identify possible leverage points and outliers.
3. Based on ANOVA, is there any variable that is not significant to explain pH?
4. What happens when we fit the following model:
   $pH = \beta_1 N + \beta_2 P + \beta_3 Ca + \beta_4 Mg + \beta_5 K + \beta_6 Na + \varepsilon$? Repeat analyses 1, 2 and 3.
5. Now include Block effects. What can we see?
6. Is there any significant block effect?
7. Is there any variable deviating from normality? Is there any transformation possible?
8. Is there some *normalization* that we could apply?

Variance

1. We are mostly interested in analysing the variance of a data set, but why?
2. Can we calculate the variance of a variable always with the same formula?
3. What is the difference between sample and population variances?
4. What is between group variance? And within group variance?
5. What about covariance?

Expectation

Before we start with variance, let us remember what is expectation:

$$E[X] = \sum_{X \in \Omega} X\mathbb{P}[x = X] \tag{19}$$

or

$$E[X] = \int_{\Omega} X\mathbb{P}[x = X]dx, \tag{20}$$

where X is a variable defined in $(\mathcal{F}, \Omega, \mathbb{P})$.

Expectation

For example, when $X \sim U(a, b)$, whose probability density function is $\mathbb{P}[x = X] = \dfrac{1}{b-a}$:

$$
\begin{aligned}
E[X] &= \int_a^b x\mathbb{P}[x = X]dx && (21)\\
&= \int_a^b x\left(\frac{1}{b-a}\right)dx && (22)\\
&= \left(\frac{1}{b-a}\right)\left[\frac{x^2}{2}\right]_a^b && (23)\\
&= \left(\frac{1}{b-a}\right)\frac{(b^2-a^2)}{2} && (24)\\
&= \left(\frac{1}{b-a}\right)\frac{(b-a)(b+a)}{2} && (25)\\
&= \frac{b+a}{2}. && (26)
\end{aligned}
$$

Exercise

1. What is the expectation of X when $X \sim U(3,7)$? Use the function *runif(n,a,b)*.
2. Plot a histogram for 10, 100, 1000, 10000 and 100000 values of X. Use the function *hist()*.
3. What is the mean of 4, 50 and 100 values generated from $U(3,7)$? Is the mean the expected value?
4. Plot the function f, such that: $f(n) = E[X|\#\Omega = n]$. What can you observe?
5. Maurício: Exponential, Ailton: Binomial, Thalita: Gamma, Gabriel: Weibull, Fernanda: Normal.

Some distributions

Table: Expectation and Variance for some distributions.

Definition	**Expected value**	**Variance**
$N(\mu, \sigma^2)$	μ	σ^2
$U(a, b)$	$\frac{b+a}{2}$	$\frac{(b-a)^2}{12}$
$Exp(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$Poisson(\lambda)$	λ	λ
$Gamma(\alpha, \beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
$Weibull(\alpha, \beta)$	$\alpha\Gamma(1 + 1/\beta)$	$\alpha^2 \left(\Gamma(1+2/\beta) - (\Gamma(1+1/\beta))^2\right)$
$Binomial(n, p)$	np	$np(1-p)$

Exercise

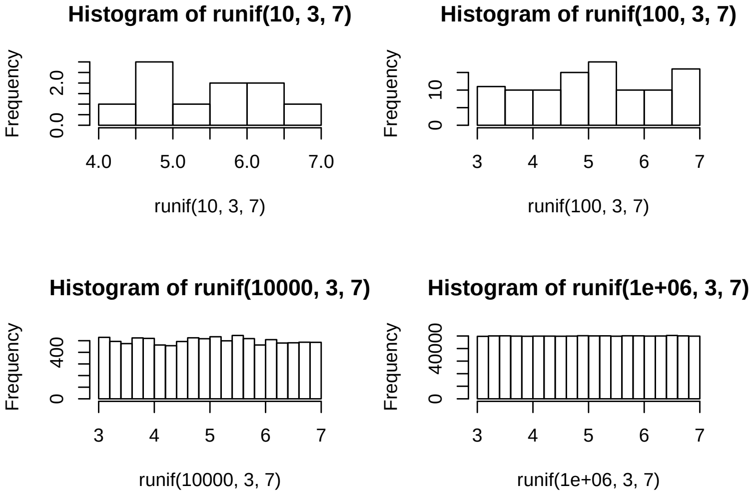

Figure: Histogram for $U(3,7)$ for different number of generated points.

Exercise

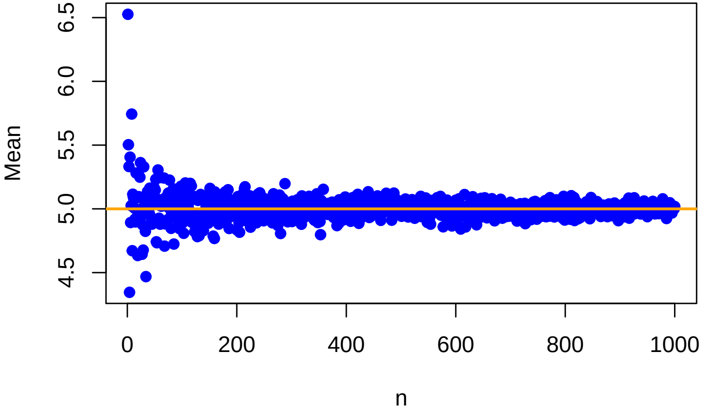

Figure: Mean of $U(3, 7)$ for different values of n. In orange, the expected value.

Exercise

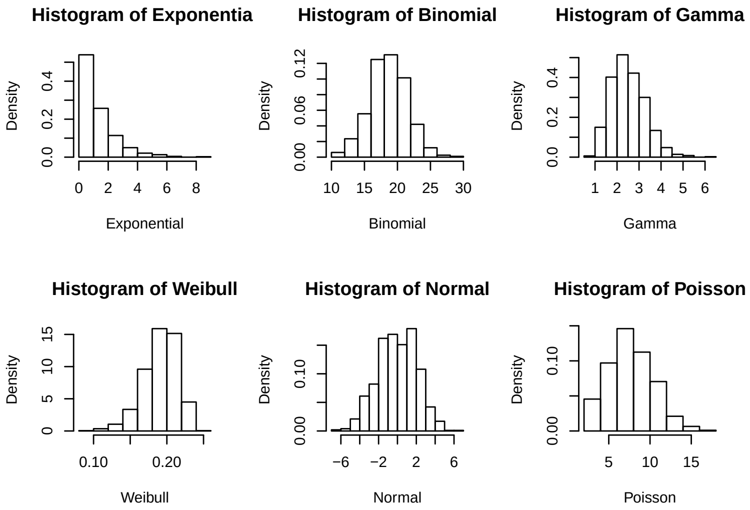

Figure: Histogram for different distributions of probability.

Exercise

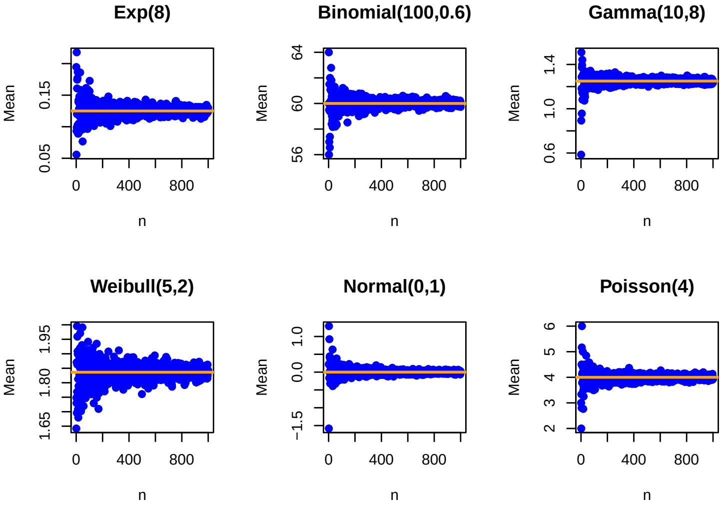

Figure: Mean of different distributions for different values of n. In organge, the expected value