Linear models

1. | Introduction

Sample coeffcients

Definition 1 (Sample variance and covariance). Let $\{(x_i, y_i) : i = 1, \dots, n\}$ be a set of data. We define the *sample covariance* between the x_i and the y_i as:

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

We define the *sample variance* as:

$$s_x{}^2 := \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Definition 2. Let $\{(x_i, y_i) : i = 1, \dots, n\}$ be a set of data. We define the *sample correlation coefficient* between the x_i and the y_i as:

$$r := \frac{s_{xy}}{s_x s_y}$$

Proposition 3. Let $\{(x_i, y_i) : i = 1, \dots, n\}$ be a set of data. Then, $r^2 \leq 1$.

Multivariate properties

Definition 4. Let $\mathbf{x} = (X_1, \dots, X_n)$ be a random vector. We define its *expectation* as:

$$\mathbb{E}(\mathbf{x}) := \begin{pmatrix} \mathbb{E}(X_1) \\ \vdots \\ \mathbb{E}(X_n) \end{pmatrix}$$

Analogously, the expectation of a matrix is defined component by component.

Theorem 5. Let $\mathbf{x}$, $\mathbf{y}$ be random vectors of dimension n. We have the following properties regarding the expectation:

1. $\mathbb{E}(\alpha\mathbf{x}+\beta\mathbf{y}+\gamma\mathbf{1}) = \alpha\mathbb{E}(\mathbf{x})+\beta\mathbb{E}(\mathbf{y})+\gamma\mathbf{1}\ \forall\alpha,\beta,\gamma \in \mathbb{R}$[1].
2. $\mathbb{E}(\mathbf{a}^{\mathrm{T}}\mathbf{x} + \mathbf{b}^{\mathrm{T}}\mathbf{y} + \mathbf{c}^{\mathrm{T}}\mathbf{1}) = \mathbf{a}^{\mathrm{T}}\mathbb{E}(\mathbf{x}) + \mathbf{b}^{\mathrm{T}}\mathbb{E}(\mathbf{y}) + \mathbf{c}^{\mathrm{T}}\mathbf{1}\ \forall\mathbf{a},\mathbf{b},\mathbf{c} \in \mathbb{R}^n$.
3. $\mathbb{E}(\mathbf{A}\mathbf{x}) = \mathbf{A}\mathbb{E}(\mathbf{x})\ \forall\mathbf{A} \in \mathcal{M}_{m\times n}(\mathbb{R})$.

Definition 6. Let $\mathbf{x} = (X_1, \dots, X_n)$ be a random vector. We define the *covariance matrix* of $\mathbf{x}$ as the following matrix:

$$\mathbf{\Sigma_x} := \mathrm{Var}(\mathbf{x}) := \mathbb{E}\left((\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{x} - \mathbb{E}(\mathbf{x}))^{\mathrm{T}}\right)$$

Proposition 7. Let $\mathbf{x} = (X_1, \dots, X_n)$ be a random vector. Then, $\mathbf{\Sigma_x}$ is symmetric and:

$$\mathbf{\Sigma_x} = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{pmatrix}$$

Definition 8. Let $\mathbf{x}$, $\mathbf{y}$ be random vectors. We define the *covariance* between them as the following matrix:

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) := \mathbb{E}\left((\mathbf{x} - \mathbb{E}(\mathbf{x}))(\mathbf{y} - \mathbb{E}(\mathbf{y}))^{\mathrm{T}}\right)$$

Proposition 9. Let $\mathbf{x}$, $\mathbf{y}$ be random vectors, $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ and $\mathbf{A}, \mathbf{B} \in \mathcal{M}_n(\mathbb{R})$. Then:

1. $\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathbb{E}(\mathbf{x}\mathbf{y}^{\mathrm{T}}) - \mathbb{E}(\mathbf{x})\mathbb{E}(\mathbf{x})^{\mathrm{T}}$
2. $\mathrm{Cov}(\mathbf{x} - \mathbf{a}, \mathbf{y} - \mathbf{b}) = \mathrm{Cov}(\mathbf{x}, \mathbf{y})$
3. $\mathrm{Cov}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) = \mathbf{A}\mathrm{Cov}(\mathbf{x}, \mathbf{y})\mathbf{B}^{\mathrm{T}}$
4. $\mathbb{E}\left(\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}\right) = \mathrm{tr}(\mathbf{A}\mathbf{\Sigma_x}) + \mathbb{E}(\mathbf{x})^{\mathrm{T}}\mathbf{A}\mathbb{E}(\mathbf{x})$
5. $\mathbb{E}\left((\mathbf{x} - \mathbf{a})(\mathbf{x} - \mathbf{a})^{\mathrm{T}}\right) = \mathbf{\Sigma_x} + (\mathbb{E}(\mathbf{x}) - \mathbf{a})(\mathbb{E}(\mathbf{x}) - \mathbf{a})^{\mathrm{T}}$
6. $\mathbb{E}(\|\mathbf{x} - \mathbf{a}\|) = \mathrm{tr}\,\mathbf{\Sigma_x} + \|\mathbb{E}(\mathbf{x}) - \mathbf{a}\|$
7. $\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma_x}\mathbf{a} = \mathrm{Var}(\mathbf{a}^{\mathrm{T}}\mathbf{x})$. Thus, $\mathbf{\Sigma_x}$ is positive semi-definite.
8. If $\mathbf{x} = (X_1, \dots, X_n)$ and no Y_j can be expressed as a linear combination of the other ones, then $\mathbf{\Sigma_x}$ is positive definite.

Multivariate normal

Definition 10. We say that $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ is symmetric and positive definite, if its moment generating function[2] is:

$$\psi_{\mathbf{x}}(\mathbf{u}) = \mathrm{e}^{\boldsymbol{\mu}^{\mathrm{T}}\mathbf{u}}\mathrm{e}^{\frac{1}{2}\mathbf{u}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{u}} \qquad \forall\mathbf{u} \in \mathbb{R}^n$$

Proposition 11. Let $\mathbf{z} = (Z_1, \dots, Z_n)$, where $Z_i \sim N(0, 1)$ for $i = 1, \dots, n$. Then, $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$.

Definition 12. Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be symmetric and positive definite. Suppose the Jordan descomposition of $\mathbf{A}$ is $\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$. Then $\forall\alpha \in \mathbb{R}$,

$$\mathbf{A}^{\alpha} := \mathbf{P}\mathbf{\Lambda}^{\alpha}\mathbf{P}^{-1}$$

where $\mathbf{\Lambda}^{\alpha} := \mathrm{diag}({\lambda_1}^{\alpha}, \dots, {\lambda_n}^{\alpha})$.

Proposition 13. Let $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathbf{\Sigma} \in \mathcal{M}_n(\mathbb{R})$ is symmetric and positive definite[3]. Then, $\mathbf{z} := \mathbf{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}) \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Analogously if $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$, then $\mathbf{x} := \mathbf{\Sigma}^{1/2}\mathbf{z} + \boldsymbol{\mu} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$.

Proposition 14. Let $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$. Then,

$$f_{\mathbf{x}}(\mathbf{y}) = \frac{1}{\sqrt{\det\mathbf{\Sigma}}}\frac{1}{(2\pi)^{\frac{n}{2}}}\mathrm{e}^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

Proposition 15. Let $\mathbf{x} \sim N_n(\boldsymbol{\mu}, \mathbf{\Sigma})$ such that $\mathbf{x} = (X_1, \dots, X_n)$. Then, the variables X_i are normal for $i = 1, \dots, n$.

[1] Here $\mathbf{1}$ represents the vector $\mathbf{1} := (1, \dots, 1)^{\mathrm{T}}$.
[2] Remember definition **??**.
[3] From now on this hypothesis will be implicit in the definition of $\mathbf{x}$.

2. | Simple regression

The model and estimations of the coefficients

Definition 16 (Simple model). Suppose we have a sample of data $\{(x_i, y_i) : i = 1, \dots, n\}$. We can describe the relationship between x_i and y_i with the following model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \dots, n$$

where we assume that y_i and ε_i are random variables whereas x_i are known constants. Moreover in the model, we suppose the following hypothesis:

1. $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, n$.
2. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if for all $i \neq j$.

Sometimes we add an additional condition of normality:

1. $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$.

If we write $\mathbf{y} = (y_1, \dots, y_n)$, $\mathbf{x} = (x_1, \dots, x_n)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ we can write the model in a more compact way:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon} \tag{1}$$

From here, we would like to estimate the parameters β_0 and β_1 to make *preditions* $\hat{y}_h$ from new data x_h.

Proposition 17 (Least-squares method). Given the simple linear model of Eq. (1), we need to estimate the parameters β_0, β_1 and σ^2. To do so, *least-squares method* seek estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of square of the deviations $y_i - \hat{y}_i$ (also called *residuals*), where $\hat{y}_i$ is the predicted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Hence:

$$\hat{\beta}_0 = \arg\min_{\beta_0} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) : \beta_0, \beta_1 \in \mathbb{R}^2 \right\}$$
$$\hat{\beta}_1 = \arg\min_{\beta_1} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) : \beta_0, \beta_1 \in \mathbb{R}^2 \right\}$$

And we obtain:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{xy}}{{s_x}^2}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

To estimate σ^2 we use:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Theorem 18. Given the model of Eq. (1), if we consider the hypothesis of normality for ε_i, then the estimates of the least-squares method coincide with the MLEs.

3. | Multiple regression

Definition 19 (General linear model). Suppose we have a sample of data $\{(x_{i1}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$. We can describe the relationship between $x_{i1}, \dots, x_{ik}$ and y_i with the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i \qquad i = 1, \dots, n$$

where we assume that y_i and ε_i are random variables whereas x_i are known constants. Moreover in the model, we suppose the following hypothesis:

1. $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, \dots, n$.
2. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ if for all $i \neq j$.

Sometimes we add an additional condition of normality:

1. $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$.

Analogously to what we did with the simple model, we can write the relation in matrix notation as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & & x_{2k} \\ \vdots & & \ddots & \vdots \\ 1 & \cdots & x_{n(k-1)} & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$
$$= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

where the matrix $\mathbf{X}$ is called *design matrix* (and its components, *regressor coefficients*), and $\boldsymbol{\beta}$, *regression coefficients*. From here, we would like to estimate the parameters $(\beta_0, \dots, \beta_k)$ with estimators $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_k)$ to make *preditions* $\hat{y}_h$ from new data $\mathbf{x}_h$ in the following way:

$$\hat{y}_h = {\mathbf{x}_h}^{\mathrm{T}} \hat{\boldsymbol{\beta}}$$

Least-squares estimation

Proposition 20 (Least-squares method). Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ of Eq. (2). We want to minimize the value

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} - y_i)^2$$

The value $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \dots, \hat{\beta}_k)$ that minimizes the previous values is given by the solution of:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

In particular, if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible, we get the explicit solution

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Proposition 21. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ of Eq. (2). If $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, then $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for $\boldsymbol{\beta}$.

Proposition 22. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$ of Eq. (2). If $\mathrm{Var}(\mathbf{y}) = \sigma^2 \mathbf{I}_n$, then the covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2 \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$.

MLE estimation

Proposition 23 (MLE method). Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2). We want to find the value $\hat{\boldsymbol{\beta}}$ that maximises the likelihood which in this case is:

$$L(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \frac{1}{(\sqrt{2\pi\sigma^2})^n} \mathrm{e}^{-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|}$$

Solving for $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ we get:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$
$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{e}\|^2 = \frac{\mathrm{SSE}}{n}$$

where $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Note that $\hat{\sigma}^2$ is biased and if we want an unbiased estimator we should use:

$$s^2 = \frac{1}{n-k-1}\|\mathbf{e}\|^2 =: \mathrm{MSE}$$

Here MSE stands for *mean square error*.

Definition 24. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) and suppose that $\mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathrm{GL}_{k+1}(\mathbb{R})$. We define the following deterministic matrices:

$$\mathbf{A} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}} \qquad \mathbf{H} = \mathbf{X}\mathbf{A} \qquad \mathbf{M} = \mathbf{I} - \mathbf{H}$$

Hence,

$$\hat{\boldsymbol{\beta}} = \mathbf{A}\mathbf{y} \qquad \hat{\mathbf{y}} = \mathbf{H}\mathbf{y} \qquad \mathbf{e} = \mathbf{M}\mathbf{y}$$

Proposition 25. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) and suppose that $\mathbf{X}^{\mathrm{T}}\mathbf{X} \in \mathrm{GL}_{k+1}(\mathbb{R})$. Then:

1. $\hat{\beta}_0, \dots, \hat{\beta}_k$ are independent normally distributed random variables, as well as $\hat{y}_1, \dots, \hat{y}_n$ and $\hat{e}_1, \dots, \hat{e}_n$
2. $\mathbf{H}$ and $\mathbf{M}$ are symmetric and idempotent. Moreover, $\mathbf{M}\mathbf{H} = \mathbf{0}_n$. Hence, they orthogonally project $\mathbb{R}^n$ into orthogonal subspaces.
3. $\mathbf{M}\mathbf{X} = \mathbf{0}$
4. $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$
5. $\operatorname{rank}\mathbf{H} = k+1$ and $\operatorname{rank}\mathbf{M} = n-(k+1)$
6. $\mathbf{X}^{\mathrm{T}}\mathbf{e} = 0$. In particular, $\sum_{i=1}^n e_i = 0$ and $\sum_{i=1}^n x_{ij}e_i = 0\ \forall j$. So the sample covariance of $(x_{1j}, \dots, x_{nj})$ and $\mathbf{e}$ is $0\ \forall j$.
7. $\hat{\mathbf{y}}^{\mathrm{T}}\mathbf{e} = 0$. Analogously, we have that the sample covariance of $\hat{\mathbf{y}}$ and $\mathbf{e}$ is 0.

Proposition 26. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2), then $\hat{\boldsymbol{\beta}}$, s^2 are unbiased estimator for $\boldsymbol{\beta}$ and σ^2, respectively, and:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Moreover, and unbiased estimator for $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \mathrm{MSE}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$$

Proposition 27. Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be symmetric and idempotent. Then, $\mathbf{A}$ is positive semi-definite. If moreover $\operatorname{rank}\mathbf{A} = r$, then $\mathbf{A}$ has r eigenvalues equal to 1 and the rest are 0.

Lemma 28. Let $\mathbf{A} \in \mathcal{M}_n(\mathbb{R})$ be symmetric and idempotent of rank d and $\mathbf{z} \sim N_n(\mathbf{0}, \mathbf{I}_n)$. Then, $\|\mathbf{A}\mathbf{z}\|^2 \sim {\chi_d}^2$.

Proposition 29. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2). Then:

1. $\hat{\boldsymbol{\beta}} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\right)$
2. $\frac{(n-k-1)s^2}{\sigma^2} \sim {\chi_{n-k-1}}^2$
3. $\hat{\boldsymbol{\beta}}$ and s^2 are independent.

Proposition 30. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) in which σ^2 is unknown. Then, for each $j = 1, \dots, n$ we have

$$\frac{\beta_j - \hat{\beta}_j}{\sqrt{\mathrm{MSE}}} \sim t_{n-k-1}$$

Proposition 31. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2). Then, if the variables $x_{i1}, \dots, x_{ik}$ are uncorrelated $\forall i = 1, \dots, n$, then the MLE estimators $\hat{\beta}_j$ coincide with the ones of the equivalent simple models:

$$y = \beta_0 + \beta_i x_i \qquad i = 1, \dots, n$$

Theorem 32 (Gauß-Markov). Consider the model of Eq. (2)[4], then the least-squares estimators for β_j, $j = 0, 1, \dots, k$, are *BLUE* (*the Best Linear Unbiased Estimator*)[5].

Corollary 33. The predicted value $\hat{y}_h$ is invariant to a full-rank linear transformation on the x's. That is:

$$\hat{y}_z = {\mathbf{z}_h}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathbf{z}} = {\mathbf{x}_h}^{\mathrm{T}}\hat{\boldsymbol{\beta}} = \hat{y}_h$$

where $\mathbf{z}_h = \mathbf{K}^{\mathrm{T}}\mathbf{x}_h$ and $\mathbf{Z} = \mathbf{X}\mathbf{K}$ is the full-rank linear transformation.

Model in centered form

Definition 34. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2). Then, for each $i = 1, \dots, n$ we can write:

$$\tilde{y}_i = y_i - \overline{y} = \beta_1\tilde{x}_{i1} + \cdots + \beta_k\tilde{x}_{ik} + \varepsilon_i \tag{3}$$

where $\tilde{x}_{ij} = x_{ij} - \overline{x}_i$. This way we obtain a *no-intercept* linear model, which is a little bit easier to estimate. The estimation of β_0 will be:

$$\hat{\beta}_0 = \overline{y} - \overline{\mathbf{x}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_1$$

where $\hat{\boldsymbol{\beta}}_1 = (\beta_1, \dots, \beta_k)^{\mathrm{T}}$

[4] Here the normality hypothesis does not play any role.

[5] That is, they have minimum variance among all linear unbiased estimators.

Proposition 35. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) in its centered form (Eq. (3)). Then:

$$\hat{\boldsymbol{\beta}}_1 = \left(\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{y}}$$

where:

$$\tilde{\mathbf{y}} = \begin{pmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix} \quad \tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \overline{x}_1 & \cdots & x_{1k} - \overline{x}_k \\ \vdots & \ddots & \vdots \\ x_{n(k-1)} - \overline{x}_1 & \cdots & x_{nk} - \overline{x}_k \end{pmatrix}$$

And so:

$$\hat{\boldsymbol{\beta}}_1 = {\mathbf{S}_{\mathbf{X}}}^{-1}\mathbf{s}_{\mathbf{X}\mathbf{y}}$$

where:

$$(\mathbf{S}_{\mathbf{X}})_{ij} = \frac{1}{n-1}\sum_{\ell=1}^{n}(x_{\ell i} - \overline{x}_i)(x_{\ell j} - \overline{x}_j)$$

$$(\mathbf{s}_{\mathbf{X}\mathbf{y}})_i = \frac{1}{n-1}\sum_{\ell=1}^{n}(x_{\ell i} - \overline{x}_i)(y_{\ell i} - \overline{y})$$

are the respective sample covariance matrices[6].

Coefficient of determination

Definition 36. Given the model of Eq. (2), we define the *coefficient of determination* R as:

$$R^2 := \frac{\mathrm{SSR}}{\mathrm{SST}} := 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} := \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ is the *regression sum of squares* and $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the *total sum of squares*. Furthermore, we can partition SST into $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$, where SSE is the *error sum of squares*. That is:

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

Lemma 37. Consider the model of Eq. (2). Then, $R^2 \leq 1$

Proposition 38. Given the simple model of Eq. (1), we have that $R^2 = r^2$.

Definition 39. Given the model of Eq. (2), we define the *adjusted coefficient of determination* R_{adj} as:

$${R_{\mathrm{adj}}}^2 = 1 - \frac{\mathrm{MSE}}{\mathrm{MST}} := 1 - \frac{n-1}{n-k-1}\frac{\mathrm{SSE}}{\mathrm{SST}}$$

where we have defined the *total mean of squares* as $\mathrm{MST} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$.

Lemma 40. Consider the model of Eq. (2). The function ${R_{\mathrm{adj}}}^2(k)$ attains a maximum at $k = k_0$ which is the optimal number of variables we should consider for our model.

Analysis of variance

Definition 41. Let $X_1 \sim \chi_{d_1}$ and $X_2 \sim \chi_{d_2}$ be independent random variables. We define the *F-distribution with degrees of freedom d_1 and d_2* as the distribution of:

$$F = \frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}$$

Proposition 42. Let $X \sim t_n$ be a random variable. Then:

$$X^2 \sim F_{1,n}$$

Lemma 43. Let $\mathbf{x}^{\mathrm{T}} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Suppose that $\sum_{i=1}^{n} {x_i}^2 = Q_1 + \cdots + Q_k$, where $Q_j = \mathbf{x}^{\mathrm{T}}\mathbf{A}_j\mathbf{x}$ is a quadratic form and $\mathbf{A}_j$ is a symmetric positive semi-definite matrix of rank r_j, $j = 1, \ldots, k$. If $r_1 + \cdots + r_k = n$, then there exists an orthogonal matrix $\mathbf{C} \in \mathcal{M}_n(\mathbb{R})$ such that if $\mathbf{y} = \mathbf{C}^{\mathrm{T}}\mathbf{x}$, then:

$$\begin{aligned} Q_1 &= {y_1}^2 + \cdots + {y_{r_1}}^2 \\ Q_2 &= {y_{r_1+1}}^2 + \cdots + {y_{r_1+r_2}}^2 \\ &\vdots \\ Q_k &= {y_{r_1+\cdots+r_{k-1}+1}}^2 + \cdots + {y_n}^2 \end{aligned}$$

Theorem 44 (Cochran's theorem). Let $\mathbf{x} = (X_1, \ldots, X_n) \in N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$. Suppose that $\sum_{i=1}^{n} {X_i}^2 = Q_1 + \cdots + Q_k$, where $Q_j = \mathbf{x}^{\mathrm{T}}\mathbf{A}_j\mathbf{x}$ is a quadratic form and $\mathbf{A}_j$ is a symmetric positive semi-definite matrix of rank r_j, $j = 1, \ldots, k$. If $r_1 + \cdots + r_k = n$, then:

1. $Q_1, \ldots, Q_k$ are independent random variables.
2. $\frac{Q_j}{\sigma^2} \sim \chi_{r_j}$, $j = 1, \ldots, k$.

Hypothesis testing

Proposition 45 (Bonferroni's method). Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) and suppose we want a confidence $k+1$-dimensional interval I for $\boldsymbol{\beta}$ of confidence $1-\alpha$. Then, it suffices to take any interval I_j for each β_j of confidence $1 - \frac{\alpha}{k+1}$ and let $I = I_0 \times \cdots \times I_k$.

Theorem 46. Let $\mathbf{y} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ be a random variable. Then:

$$\frac{\mathrm{SSE}}{\sigma^2} \sim {\chi_{n-k-1}}^2$$

Moreover if $\boldsymbol{\beta}_1 := (\beta_1, \ldots, \beta_k)^{\mathrm{T}} = 0$ we have:

$$\frac{\mathrm{SSR}}{\sigma^2} \sim {\chi_k}^2$$

Hence in this case we have that:

$$\frac{\mathrm{SSR}/k}{\mathrm{SSE}/(n-k-1)} \sim F_{k,n-k-1}$$

Proposition 47. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2) and suppose we want to test the hypothesis $\mathcal{H}_0 : \boldsymbol{\beta} = \mathbf{0}$ versus $\mathcal{H}_1 : \boldsymbol{\beta} \neq \mathbf{0}$. The statistic that we take is:

$$F = \frac{\mathrm{SSR}/k}{\mathrm{SSE}/(n-k-1)} \sim F_{k,n-k-1}$$

[6] Note that the expression for $\hat{\boldsymbol{\beta}}_1$ is quite similar to the least-square estimate for β_1 in the simple linear model.

Proposition 48. Consider the model $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ of Eq. (2), $\mathbf{R} \in \mathcal{M}_{r\times k+1}(\mathbb{R})$ where $r \leq k+1$, $\mathbf{c} \in \mathbb{R}^{k+1}$ and suppose we want to test the hypothesis $\mathcal{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{c}$ versus $\mathcal{H}_1 : \mathbf{R}\boldsymbol{\beta} \neq \mathbf{c}$. Without loss of generality suppose we can eliminate the first $\beta, \ldots, \beta_r$ under $\mathcal{H}_0$ and so rearranging the model equation we obtain:

$$\mathbf{y}_\ell = \mathbf{X}_\ell\boldsymbol{\beta}_\ell + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}_\ell = (\beta_{r+1}, \ldots, \beta_k)$. The statistic that we take is:

$$F = \frac{(\mathrm{SSE}_\ell - \mathrm{SSE})/r}{\mathrm{SSE}\,/(n-k-1)} = \frac{(\mathrm{SSR} - \mathrm{SSR}_\ell)/r}{\mathrm{SSE}\,/(n-k-1)} \sim F_{r,n-k-1} \tag{4}$$

where the subindex in SSE_ℓ and SSR_ℓ indicates that this is the SSE and SSR for the model under $\mathcal{H}_0$, respectively.

Dummy variables

Definition 49 (Model without interaction). Suppose we have a linear model of the form of Eq. (2). We would like to measure the change in the response when adding a binary parameter (say men and women) to the model. Hence we identify that deterministic variable with a *dummy variable* as follows:

$$d := \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases} \tag{5}$$

And we consider the new model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + d\gamma\mathbf{1} + \boldsymbol{\varepsilon}$$

where $\mathbf{1} = (1, \ldots, 1)$. Finally we may want to conclude whether the variable d is relevant or not. This can be obtained by doing the t-test $\mathcal{H}_0 : \gamma = 0$[7]. Note that, independently of the conclusion of the test, both regression lines (the one for men and the one for women) will be parallel.

Definition 50 (Model with interaction). Suppose we have a linear model of the form of Eq. (2). We would like to measure the change in the response when adding a dummy variable d as in Eq. (5). Now, consider the following *model with interaction*:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + d\mathbf{X}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

Let $\gamma := \delta_0$. And now we can do tests for $\boldsymbol{\delta}$ or for γ. Note that, these regression lines won't be in general parallel.

Definition 51 (Segmented regression). Suppose we have a linear model of the form of Eq. (2) and that for some reason we have noticeable change of the slope at $\mathbf{x}^*$ (known). Then, defining the dummy variable

$$d_i := \begin{cases} 1 & \text{if } x_{ki} \geq x_i{}^* \\ 0 & \text{if } x_{ki} \leq x_i{}^* \end{cases} \tag{6}$$

we can consider the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Y}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where:

$$\mathbf{Y} = \begin{pmatrix} d_1(x_{11} - x_1{}^*) & \cdots & d_k(x_{1k} - x_k{}^*) \\ \vdots & \ddots & \vdots \\ d_1(x_{n1} - x_1{}^*) & \cdots & d_k(x_{nk} - x_k{}^*) \end{pmatrix}$$

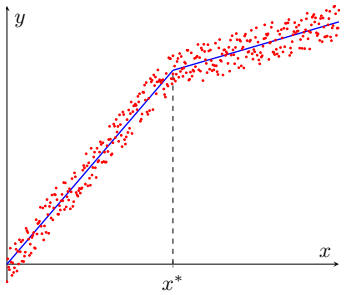


Figure 1: Segmented regression of a simple linear model

Predicted confidence intervals

Definition 52. Consider the model of Eq. (2) and suppose we have observed a new data value $\mathbf{x}_h{}^{\mathrm{T}} = (1, x_{1h}, \ldots, x_{kn})$. The *response* or *prediction* of this observation is

$$y_h = \mathbf{x}_h{}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_h = \mu_h + \varepsilon_h$$

where μ_h is the *average response*. The *estimated average response* $\hat{\mu}_h$ and the *estimation of the prediction* $\hat{y}_h$ coincide and they are:

$$\hat{y}_h = \hat{\mu}_h = \mathbf{x}_h{}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$$

Proposition 53 (Confidence interval for the average response). Consider the model of Eq. (2) and suppose we have observed a new data value $\mathbf{x}_h{}^{\mathrm{T}} = (1, x_{1h}, \ldots, x_{kn})$. Then:

1. The random variable $\hat{\mu}_h$ is a linear combination of $y_1, \ldots, y_n$ and:

   $$\mathbb{E}(\hat{\mu}_h) = \mu_h \qquad \mathrm{Var}(\hat{\mu}_h) = \sigma^2 h_{hh}$$

   where we have defined the *leverage* as $h_{hh} := \mathbf{x}_h{}^{\mathrm{T}}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{x}_h$.

2. $\hat{\mu}_h \sim N(\mu_h, \sigma^2 h_{hh})$

3. $\dfrac{\hat{\mu}_h - \mu_h}{\sqrt{\mathrm{MSE}\, h_{hh}}} \sim t_{n-k-1}$

Varying the value of $\mathbf{x}_h$ along the space of parameters we obtain a family of confidence intervals for y_h which is called *confidence band*.

Proposition 54 (Confidence interval for the predicted value). Consider the model of Eq. (2) and suppose we have observed a new data value $\mathbf{x}_h{}^{\mathrm{T}} = (1, x_{1h}, \ldots, x_{kn})$. Then:

1. The random variable $y_h - \hat{y}_h$ is a linear combination of $y_1, \ldots, y_n$ and:

   $$\mathbb{E}(y_h - \hat{y}_h) = 0 \qquad \mathrm{Var}(y_h - \hat{y}_h) = \sigma^2(1 + h_{hh})$$

[7] Note that this test is equivalent to the one in Eq. (4) due to Theorem 42.

2. $y_h - \hat{y}_h \sim N(0, \sigma^2(1 + h_{hh}))$

3. $$\frac{y_h - \hat{y}_h}{\sqrt{\text{MSE}(1 + h_{hh})}} \sim t_{n-k-1}$$

Varying the value of $\mathbf{x}_h$ along the space of parameters we obtain a family of confidence intervals for y_h which is called *prediction band*.

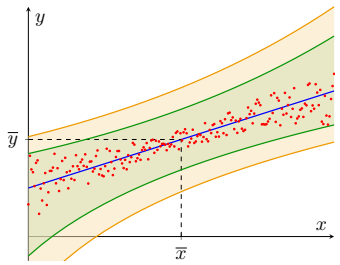


Figure 2: Example of a confidence band (green-shaded region) and a prediction band (yellow-shaded region) for a simple linear model. Note that always the prediction band is greater than the confidence band

Corollary 55. Consider the simple model of Eq. (1) and suppose we have observed a new data value x_h. Then:

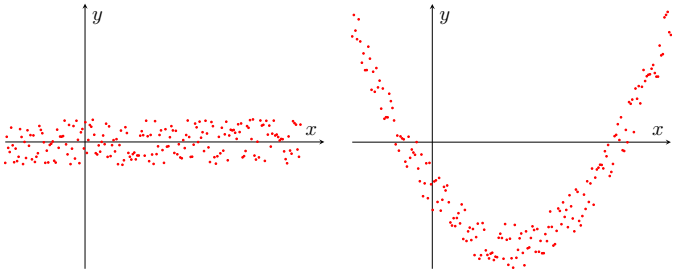
$$h_h := h_{hh} = \frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

Lack of fit

Definition 56. Consider the simple model of Eq. (1). Then, the residuals $e_i = y_i - \hat{y}_i$ satisfy:

1. $\sum_{i=1}^n e_i = 0$
2. $\sum_{i=1}^n x_i e_i = 0$
3. $\hat{\sigma}_e = \frac{\sum_{i=1}^n (e_i - \overline{e})}{n-2} = \text{MSE}$.
4. They are pairwise correlated.
5. $\mathbb{E}(e_i) = 0$
6. $\text{Var}(e_i) = \sigma^2(1 - h_i)$

Proposition 57. Consider the simple model of Eq. (1). We will perceive a lack of fit in the model if the graph of the e_i in terms of the regressors coefficients x_i and the prediction $\hat{y}_i$ follow a chaotic behavior inside of a rectangle centered at $\hat{y} = 0$ (see Fig. 3).



(a) Absence of lack of fit (b) Lack of fit

Figure 3: Stability of limit cycles

Proposition 58 (LOF test). Consider the simple model of Eq. (1). The *lack of fit test* (*LOF test*) has a the following null hypothesis:

$\mathcal{H}_0$: For each x_i, the mean of all outcomes obtained for this fixed value lies on the regression line

Otherwise, $\mathcal{H}_1$: There is a LOF. Suppose now that we have m distinct x_i and for each of those we have the observation y_{ij}, $j = 1, \ldots, n_i$. In order to properly do this test we need to have at least one index $i = i^*$ such that number $n_{i^*} \geq 2$. In this case, we have that:

$$\text{SSE} = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \hat{y}_i)^2 = \sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 + \\ + \sum_{i=1}^m n_i (\overline{y}_i - \hat{y}_i)^2 =: \text{SSPE} + \text{SSLOF}$$

where $\overline{y}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$. Finally the test that we take is (under $\mathcal{H}_0$):

$$\frac{\frac{\text{SSLOF}}{m-2}}{\frac{\text{SSPE}}{n-m}} =: \frac{\text{MSLOF}}{\text{MSPE}} \sim F_{m-2, n-m}$$

Definition 59. Consider the simple model of Eq. (1). In order to normalize the errors we define the *internally studentized residuals* as:

$$r_i = \frac{e_i}{\sqrt{\text{MSE}(1 - h_i)}} \sim t_{n-k-1}$$

We define the *externally studentized residuals* as:

$$dr_i = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1 - h_i)}} \sim t_{n-k}$$

where $\text{MSE}_{(i)} = \frac{1}{n-k} \sum_{\substack{j=1 \\ j \neq i}}^n (y_j - \hat{y}_j)^2$.

Definition 60. Consider the simple model of Eq. (1). There are two types of atypical data: the *high-leverage points* and the *outliars*. To detect them, we will say the the data i is a *high-leverage point* if:

$$h_{ii} \geq 3\overline{h} = 3\frac{k+1}{n}$$

We will say that the data j such that $|dr_j| = \max |dr_i| : i = 1, \ldots, n$ is an outliars if

$$2n\gamma \geq \alpha = 0.05$$

where $\gamma = 1 - \mathbb{P}(t_{n-k} > |dr_j|)$.

Definition 61. Consider the simple model of Eq. (1). We will say that the influence of the i-th point on $\hat{\beta}_j$ is significative if

$$\left|\text{DFBETAS}_{j(i)}\right| := \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\text{MSE}_{(i)}\, b_{jj}}} \geq \frac{2}{\sqrt{n}}$$

where $\hat{\beta}_{j(i)}$ and $\text{MSE}_{(i)}$ are the estimator of $\hat{\beta}_j$ and MSE in a model without the i-th point and $b_i i$ is the i-th element on the diagonal of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

Definition 62. Consider the simple model of Eq. (1). We will say that the influence of the i-th point on the prediction is significative if

$$\left|\text{DFFITS}_{j(i)}\right| := \frac{\hat{y}_j - \hat{y}_{i(i)}}{\sqrt{\text{MSE}_{(i)}\, h_i i}} \geq 2\sqrt{\frac{p}{n}}$$

where $\hat{y}_{i(i)}$ is the prediction $\hat{y}_i$ in a model without the i-th point.

Definition 63. Consider the simple model of Eq. (1). We will say that the i-th point has a global influence the influence on the model if the *Cook's distance*

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(k+1)\,\text{MSE}} = \frac{{r_i}^2}{k+1}\frac{h_{ii}}{1-h_{ii}}$$

satisfy:

$$D_i \geq 1$$

Multicollinearity

Definition 64. Consider the simple model of Eq. (2). We will have *multicollinearity* in our data if $\exists \mathbf{c} \in \mathbb{R}^{k+1}$ such that $\mathbf{Xc} \simeq \mathbf{0}$. Thus, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ will be approximately singular. In particular, since $\mathbf{\Sigma}_{\hat{\boldsymbol{\beta}}} = \sigma^2\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}$, we will notice a large variance on the approximations.

Definition 65. Consider the simple model of Eq. (2). We define the *variance inflation factor* (or *VIF*) of the data $\mathbf{x}_j = (x_{1j}, \ldots, x_{n1})^{\mathrm{T}}$ as:

$$\text{VIF}(\mathbf{x}_j) := \frac{1}{1-{R_j}^2}$$

where R_j is the coefficient of determination of the regression of $\mathbf{x}_j$ in terms of the data $\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_k$.

Lemma 66. Consider the simple model of Eq. (2). Then:

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{(n-1){s_j}^2}\,\text{VIF}(\mathbf{x}_j)$$

where ${s_j}^2 := \text{Var}(\mathbf{x}_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2$.

Definition 67. Consider the simple model of Eq. (2). We define the *tolerance* of multicollinearity as the inverse of VIF.

Definition 68. Consider the simple model of Eq. (2). We impose that the data $\mathbf{x}_j$ is affected by multicollinearity if $\text{VIF}(\mathbf{x}_j) \geq 5$. Or alternatively, if the tolerance if ≤ 0.2. Thus, we will proceed to remove it unless it is significant for the model.

Theorem 69. Let $\mathbf{A}, \mathbf{B} \in \mathcal{M}_k(\mathbb{R})$ be symmetric matrices such that $\mathbf{A}$ is positive semi-definite. and $\mathbf{B}$ is positive definite. Let $Q(\mathbf{v}) = \mathbf{v}^{\mathrm{T}}\mathbf{A}\mathbf{v}$, $\mathbf{v} \in \mathbb{R}^k$. Then:

$$\lambda_1 = \max\{Q(\mathbf{v}) : \mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v} = 1\}$$
$$\mathbf{v}_1 = \arg\max\{Q(\mathbf{v}) : \mathbf{v}^{\mathrm{T}}\mathbf{B}\mathbf{v} = 1\}$$

where $\mathbf{v}_1$ is the eigenvector of the largest eigenvalue λ_1 of $\mathbf{B}^{-1}\mathbf{A}$.

Mallow's C_p statistic

Definition 70. Consider the model of Eq. (2) and let $p - 1 < k$. We would like to compare the original model with the model when the variables $\mathbf{x}_j$, $j = p, \ldots, k$, are removed. We define the following matrices:

$$\mathbf{X}_1 = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & & x_{2(p-1)} \\ \vdots & & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n(p-1)} \end{pmatrix}$$
$$\mathbf{X}_2 = \begin{pmatrix} x_{1p} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{np} & \cdots & x_{nk} \end{pmatrix}$$

That is, $\mathbf{X} = (\mathbf{X}_1 \mid \mathbf{X}_2)$, expressed as a block matrix. Similarly we define $\boldsymbol{\beta}_1 = (\beta_0, \ldots, \beta_{p-1})^{\mathrm{T}}$ and $\boldsymbol{\beta}_2 = (\beta_p, \ldots, \beta_k)^{\mathrm{T}}$. Finally, we define $\hat{\boldsymbol{\beta}}_1$ to be the estimators of the new model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1$. That is:

$$\hat{\boldsymbol{\beta}} = \left({\mathbf{X}_1}^{\mathrm{T}}\mathbf{X}_1\right)^{-1}{\mathbf{X}_1}^{\mathrm{T}}\mathbf{y}$$

Proposition 71 (Bias on the estimations). Consider the model of Eq. (2) and the new model considering only the first $p-1 < k$ columns of data. Then, the bias of each the new estimation $\hat{\boldsymbol{\beta}}_1$ is:

$$\text{bias}(\hat{\boldsymbol{\beta}}_1) = \left({\mathbf{X}_1}^{\mathrm{T}}\mathbf{X}_1\right)^{-1}{\mathbf{X}_1}^{\mathrm{T}}\mathbf{X}_2{\boldsymbol{\beta}_2}^{[8]}$$

Proposition 72 (Bias on the predictions). Consider the model of Eq. (2) and the new model considering only the first $p-1 < k$ columns of data. Then, the bias of each the predictions $\hat{\mathbf{y}}_1$ is:

$$\text{bias}(\hat{\mathbf{y}}_1)^2 = \mathbf{X}\boldsymbol{\beta}^{\mathrm{T}}(1 - \mathbf{H}_1)\mathbf{X}\boldsymbol{\beta}$$

where we have defined $\mathbf{H}_1 := \mathbf{X}_1\left({\mathbf{X}_1}^{\mathrm{T}}\mathbf{X}_1\right)^{-1}{\mathbf{X}_1}^{\mathrm{T}}$.

Theorem 73 (Mallow's C_p statistic). Consider the model of Eq. (2) and the new model considering only the first $p-1 < k$ columns of data. We define the *Mallow's C_p statistic* as:

$$\text{C}_p := \frac{\text{SSE}_p}{\text{MSE}} - (n - 2p)$$

where SSE_p is the error sum of squares of the new model with p variables. From here, we can conclude that the closer C_p is to p (i.e. $\text{C}_p \approx p$) the lower the bias is for this new model. Moreover the smaller C_p is, the lower the mean square error is.

Remark. Given our initial data for the model of Eq. (2), we can get some submodels by considering the best ones regarding the residuals, multicollinearity, C_p statistic... Once we have all these submodels $M_1, \ldots, M_r$ done, we should choose only one of them among the others. To do so, we randomly partitionate our data in ℓ-folds to obtain deterministic matrices $\mathbf{X}_1, \ldots, \mathbf{X}_\ell$ (each with $k+1$ columns) and vectors $\mathbf{y}_1, \ldots, \mathbf{y}_\ell$. Now, for each $j \in$

[8] Note that if the submatrices $\mathbf{X}_1$ and $\mathbf{X}_2$ are orthogonal we won't have bias.

$\{1, \dots, \ell\}$ and for each model M_i we measure how good is the prediction of $\mathbf{y}_j$ (with the SSE) taking all the other data. We obtain thus a value $\text{SSE}_{i,j}$ for each i, j. We will take the the model i_0 such that:

$$i_0 = \underset{i \in \{1,\dots,r\}}{\arg\min} \sum_{j=1}^{\ell} \text{SSE}_{i,j}$$

Information and entropy

Axiom 74 (Information). The *information of an event* in a probability space is a continuous function $I(A)$ that satisfies the following properties:

- It depends on $\mathbb{P}(A)$.
- It increases as $\mathbb{P}(A)$ decreases and if $\mathbb{P}(A) = 1$, then $I(A) = 0$.
- $I(A \cap B) = I(A) + I(B)$ for any event A, B.

Theorem 75 (Cauchy's functional equation). Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function such that

$$f(x + y) = f(x) + f(y)$$

Then, $\exists c \in \mathbb{R}$ such that $f(x) = cx \ \forall x \in \mathbb{R}$.

Proof. We will prove that $f(q) = cq \ \forall q \in \mathbb{Q}$ and the density of $\mathbb{Q}$ in $\mathbb{R}$ will finish the proof.
First note that $f(0) = f(0 + 0) = 2f(0)$, so $f(0) = 0$ and therefore $0 = f(x - x) = f(x) + f(-x)$, which implies that the function f is odd. Let $n \in \mathbb{N}$ and $x \in \mathbb{R}$. Then:

$$f(nx) = f((n-1)x) + f(x) = \cdots = nf(x)$$

If $x = \frac{1}{n}$, then $f(\frac{1}{n}) = f(1)\frac{1}{n}$ and define $c := f(1)$. Finally, $\forall \frac{n}{m} \in \mathbb{Q}$ we have:

$$f\left(\frac{n}{m}\right) = nf\left(\frac{1}{m}\right) = c\frac{n}{m}$$

□

Lemma 76. The information of an event A is

$$I(A) = -c \log \mathbb{P}(A)$$

for some $c \in \mathbb{R}_{>0}$[9].

Proof. By the first axiom of 74 Information we have that $I(A) = f(\mathbb{P}(A))$ and by the third one that if A, B are independent we have that:

$$f(\mathbb{P}(A)\mathbb{P}(B)) = f(\mathbb{P}(A \cap B)) = f(\mathbb{P}(A)) + f(\mathbb{P}(B))$$

That is, $f(xy) = f(x) + f(y) \ \forall x, y \in [0, 1]$. Now consider $g(x) = f(\mathrm{e}^x)$. Then:

$$g(x + y) = f(\mathrm{e}^x \mathrm{e}^y) = f(\mathrm{e}^x) + f(\mathrm{e}^y) = g(x) + g(y)$$

So by 75 Cauchy's functional equation, we have that $g(x) = cx$ and so $f(x) = c \log x$ for some $c \in \mathbb{R}$. The second axiom of 74 Information implies $c < 0$. □

[9] Usually we take $c = 1$ or $\frac{1}{\log b}$ in order to use the logarithm in base b.

Definition 77 (Entropy). Let X be a random variable. We define the *entropy* of X as:

$$\mathrm{H}(X) := \mathbb{E}(I(X))$$

In particular if X has outcomes in $\mathcal{X}$ and it is discrete, then:

$$\mathrm{H}(X) = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x)$$

where p is the pmf. If it is continuous we have:

$$\mathrm{H}(X) = -\int_{x \in \mathcal{X}} f(x) \log_b f(x) \, \mathrm{d}x$$

where f is the pdf.

Definition 78. Let X, Y be two random variables of support $\mathcal{X}$. We define the *Kullback-Leibler divergence* as the quantity:

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \left(\frac{p(x)}{q(x)}\right)$$

where p, q are the pmf of X and Y (in the discrete case) or as:

$$D_{\mathrm{KL}}(f \parallel g) = \int_{x \in \mathcal{X}} f(x) \log \left(\frac{f(x)}{g(x)}\right) \mathrm{d}x$$

where f, g are the pdf of X and Y (in the continuous case).

Remark. Observe that in both discrete and continuous cases we have:

$$D_{\mathrm{KL}}(p \parallel q) = \mathrm{H}(X, Y) - \mathrm{H}(Y)$$

where we have denoted $\mathrm{H}(X, Y) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$ (or $\mathrm{H}(X, Y) = -\int_{x \in \mathcal{X}} p(x) \log q(x) \, \mathrm{d}x$) the mixed entropy between p and q. Thus, $D_{\mathrm{KL}}(p \parallel q)$ expresses in some sense how different is if p from q. The lower $D_{\mathrm{KL}}(p \parallel q)$ is, the more closer is q to p.

Definition 79 (Akaike information criterion). Consider the model of Eq. (2). The *Akaike information criterion* (*AIC*) is an estimator of prediction error that is defined as:

$$\text{AIC} = 2k - 2 \log\left(\Phi_n(y; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2)\right)$$

where Φ_n is the pdf of $N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$.

Remark. The lower AIC is, the better a model is.

4. | Generalized linear models

Box-Cox transformation

Definition 80. Consider a simple model like in Eq. (1) in which we have observed a notable LOF. We define the *Box-Cox transformation* as the following transformation on the y:

$$y^{(\lambda)} := \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases}$$

Proposition 81. Consider a simple model like in Eq. (1) in which we have observed a notable LOF. Making the Box-Cox transformation and assuming that for some unknown λ the transformed observations satisfy the normal hypothesis of a linear model, we will choose the MLE $\lambda = \hat{\lambda}$ that maximises the likelihood of the problem.

Exponential families

Definition 82. Let $\{f(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ be a family of pdfs from a probability distribution. We say that this family is *exponential* if $f(\mathbf{x}, \boldsymbol{\theta})$ can be expressed as:

$$f(\mathbf{x}, \boldsymbol{\theta}) = h(x)\mathrm{e}^{\boldsymbol{\theta}^{\mathrm{T}} \cdot \mathbf{T}(\mathbf{x}) - A(\boldsymbol{\theta})} \tag{7}$$

where $\boldsymbol{\eta}$ is a function of $\boldsymbol{\theta}$.

Definition 83. Given a probability $p \in (0, 1)$ we define the *odds* as the value:

$$\frac{p}{1-p}$$

It measures how likely the event of probability p in a scale of $(0, \infty)$.

Definition 84. Given a probability $p \in (0, 1)$, we define the *log-odds* (or *logit*) as:

$$\operatorname{logit} p := \log\left(\frac{p}{1-p}\right)$$

Proposition 85. The families of Bernoulli[10], Poisson, Binomial, normal, exponential, gamma, chi-squared and beta distributions are all exponential families.

Theorem 86. Let $\{f(\mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ an exponential family that can be written as Eq. (7). Then, the statistic $\mathbf{T}$ is a sufficient statistic.

[10] In the Bernoulli's case, $\eta(p) = \operatorname{logit}(p)$.