Statistical Modelling

Ralf Blöchlinger

Contents

1 Refresher on Linear Algebra, Probability, and Statistics

1.1 Linear Algebra

1.1.1 Vectors

We confine ourselves to the real numbers.

Definition 1.1 Vector

A *vector* is an element $x \in \mathbb{R}^n$.

A *linear combination of vectors* is given by

$$b_1 \cdot x_1 + \ldots + b_p \cdot x_p$$

for some $b_i \in \mathbb{R}$ and vectors in R^n.

Definition 1.2 Span

The *span* of a nonemtpy set of vectors S consists of all linear combinations of the vectors in S.

Definition 1.3 Basis

A set of vectors S is a *basis* for a space V if

$$\text{span}(S) = V$$

We are interested in assining properties such as lengths and angles to vectors for which we introduce the scalar product.

Definition 1.4 Scalar product

The *scalar* product of two vectors $a, b \in \mathbb{R}^n$ is defined as

$$\langle a, b \rangle = b^T a = \sum_{i=1}^{n} a_i b_i$$

b^T denotes the transpose of vector b. Geometrically, the scalar product is the product of the Euclidean magnitudes of the two vectors and the cosine of the angle between them. We therefore define

Definition 1.5 Length of a vector

If $a \in \mathbb{R}^n$, then the *length* of a is defined as

$$|a|^2 = \langle a, a \rangle$$

An often useful result, which can be compared to the triangle inequality for real numbers, is the following:

Proposition 1.1 Cauchy-Schwartz Inequality

Let $a, b \in \mathbb{R}^n$. Then

$$|\langle a, b\rangle|^2 \leq \langle a, a\rangle \cdot \langle b, b\rangle$$

We give perpendicular vectors a special name. We will often encounter such vectors.

Definition 1.6 Orthogonal vectors

We call two vectors $a, b \in \mathbb{R}^n$ *orthogonal* if

$$\langle a, b\rangle = 0$$

One can see that this corresponds to the vectors being perpendicular by noting that

$$\cos\theta \cdot (\|a\|\|b\|) = a \cdot b$$

where $\|x\| = \sqrt{|\langle x, x\rangle|^2}$, and observing that $\cos\frac{\pi}{2} = 0$.

Definition 1.7 Orthonormal vectors

We call two vectors $a, b \in \mathbb{R}^n$ *orthonormal* if they are orthogonal and each vector has unit length.

A useful result, whose proof follows almost 'by inspection', is the following:

Proposition 1.2

If $e_1, \ldots, e_n$ are an orthgonal basis of R^n and

$$y = b_1 \cdot e_1 + \ldots + b_n \cdot e_n$$

then $\langle y, e_1\rangle = b_1$.

1.1.2 Matrices

A *matrix* is a rectangular array of numbers, symbols, or expressions, arranged in rows and columns, which is used to represent a mathematical object.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = (a_{ij}) \in \mathbb{R}^{m\times n}$$

A $m \times n$ matrix consists of m rows and n columns.

Whenever we multiply a matrix X with a vector b, assuming that dimensions are conformable, we obtain a linear function

$$y = Xb$$

whereby y is a linear combination of columns of X, using weights in b.

Definition 1.8 Identitiy Matrix

The *identitiy matrix* is an $n \times n$ matrix whose diagonal elements all consists of 1 and all off-diagonal elements are 0. We denote the identity matrix by

$$I = I_{n\times n} = \mathbb{I} = \text{diag}(1)_{n\times n}$$

Definition 1.9 Transpose Matrix

The *transpose* of a matrix is a matrix 'mirrored' on the diagonal.

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}$$

Under some circumstances, a square matrix is *invertible*. An invertible matrix can *undo* the linear transformation of the vector space created by the original matrix.

Definition 1.10 Invertible Matrix and Inverse

A matrix A is **invertible**, if $\exists B$, s.t.:

$$AB = BA = I_n$$

We call B the **inverse matrix** of A and denote it by A^{-1}

A result we wil often use is the following:

Proposition 1.3

The matrix $A = (X^T X)^{-1}$ is symmetric.

1.1.3 Orthogonal projections

Let $n \geq p \in \mathbb{N}$. Consider a point $y \in \mathbb{R}^n$ and a subspace S spanned by linearly independent vectors $x_1, \ldots, x_p$. We are interested in finding the orthogonal projection of y onto S. An orthgonal projection gives us the vector in S that lies 'closest' to y in the sense that the distance between the y and all vectors of the subspace S is minimised.

Definition 1.11

Let $X \in \mathbb{R}^{n\times p}$ be the matrix with columns being the linear independent vectors $x_1, \ldots, x_p$ each of length n, spanning a vector space S. The matrix

$$P_S = X\left(X^T X\right)^{-1} X^T \tag{1.1}$$

is called the **orthogonal projection matrix** onto S.

Proposition 1.4

For any $y \in \mathbb{R}^n$ and set of linear independent vectors $x_1, \ldots, x_p$, each of length n, spanning a vector space S, the vector defined by

$$v = P_s y$$

is the vector in the space S which is closest to y.

We can think of the orthogonal projection matrix as applying two operations:

1. *Analysis:* $\left(X^T X\right)^{-1} X^T$ we find 'coordinates' or 'coefficients'
2. *Synthesis:* We use our above found coordinates, to form a linear combination of the columns in X.

We now discuss some important properties of orthogonal projections.

Proposition 1.5

The projection matrix P is idempotent and symmetric. I.e.,

$$P = P^2 = P^T$$

This proposition can be intuitively understood as follows: if we project a vector into a subspace S and then once again apply the same projection, the resulting vector will not change since it already lies in the subspace.

Proposition 1.6

Each eigenvalue of P is either 0 or 1.

Proof. If A is idempotent, λ is an eigenvalue and v a corresponding eigenvector then

$$\lambda v = Av = AAv = \lambda Av = \lambda^2 v$$

Since $v \neq 0$ we find $\lambda - \lambda^2 = \lambda(1 - \lambda) = 0$ so either $\lambda = 0$ or $\lambda = 1$.

□

Proposition 1.7

The rank of P is equal to the trace of $P = p$.

Proof. One can show that projection matrices are *diagonalizable*. I.e., there exists a diagonal matrix of eigenvalues E and matrix C such that

$$P = CEC^{-1}$$

Noting the cyclical property of the trace,

$$\mathrm{tr}(P) = \mathrm{tr}(CEC^{-1}) = \mathrm{tr}(C^{-1}CE)$$

Thus the trace of P is equal to p.

This is equal to the rank of P, since the rank is equal to the number of non-zero eigenvalues for a diagonalizable matrix. □

Proposition 1.8

The matrix $1 - P$ is also a projection and $P(1 - P) = 0$

1.1.4 Quadratic Forms and positiv definite matrices

For any $n \times n$ matrix A and vector of length n $\mathbf{x}$, the quadratic form is given by

$$\mathbf{x}^T A \mathbf{x}$$

Definition 1.12 Positive (semi-)definite Matrix

A $n \times n$ matrix A is **positive definite** if

$$x^T A x > 0 \text{ for all } x \in \mathbb{R}^n \backslash \{0\}$$

A is **positive semi-definite** if

$$x^T A x \geq 0 \text{ for all } x \in \mathbb{R}^n$$

Proposition 1.9

All eigenvalues of a positive-definite matrix A are positive.

Proof. Assume for contradiction that some eigenvalue λ of A is zero or negative. Let x be the associated eigenvector. Then

$$x^T A x = \lambda x^T x = \lambda |x|^2 \leq 0$$

which contradicts that A is PD. □

Using similar arguments relating traces and determinants to eigenvalues as above, we have the following corollary:

Corollary 1.1

The trace and determinant of a positive-definite matrix A are positive.

The following result will at times be useful:

Proposition 1.10

Let $\mathbf{B}$ be an $n \times p$-matrix. The matrix $\mathbf{B}^T\mathbf{B}$ is then symmetric and nonnegative definite. If $\text{rk}(\mathbf{B}) = p$, then $\mathbf{B}^T\mathbf{B}$ is positive definite.

One way of seeing this is to note that if x is any vector of appropriate dimensions then

$$\begin{aligned} x^T\mathbf{B}^T\mathbf{B}x &= (\mathbf{B}x)^T\mathbf{B}x \\ &= \sum(\mathbf{B}x)_{ii}^2 \geq 0 \end{aligned}$$

On occasion, we will use the following two results, relating positive-definite matrices to other matrices with nice properties:

Proposition 1.11

If A is psd, then there exists a unique symmetric square root matrix B such that

$$A = BB = BB^T$$

The proof uses the Eigendecomposition.

Proposition 1.12 Choleksy decomposition

For symmetric positive-defintie matrix A, there exists a real lower triangular matrix with positive diagonal entries L such that

$$A = LL^T$$

1.2 Probability

1.2.1 Expectations

$$\mathbb{E}(X) = \int_\Omega xf(x)$$

Properties:

- Linearity: $\mathbb{E}[a\mathbf{X} + b\mathbf{Y}] = a\mathbb{E}[\mathbf{X}] + b\mathbb{E}[\mathbf{Y}]$
- For independent components: $\mathbb{E}[X_iX_j] = \mathbb{E}[X_i]\mathbb{E}[X_j]$

Estimator:

$$\hat{\boldsymbol{\mu}} = \frac{1}{m}\sum_{i=1}^{m}\mathbf{x}^{(i)}$$

Covariance matrix:

$$\Sigma = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]$$

With elements:

$$\Sigma_{ij} = \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$$

Univariate case:

$$\mathrm{Cov}(X,Y) = \mathbb{E}\big[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))\big] = \mathbb{E}\big[X \cdot Y\big] - \mathbb{E}(X)\mathbb{E}(Y)$$

Properties:

- Symmetry: $\Sigma = \Sigma^T$
- Positive semi-definite: $\mathbf{v}^T \Sigma \mathbf{v} \geq 0$ for all vectors $\mathbf{v}$
- For linear transformation $\mathbf{Y} = A\mathbf{X}$: $\mathrm{Cov}(\mathbf{Y}) = A\Sigma A^T$

Estimator:

$$\hat{\Sigma} = \frac{1}{m-1} \sum_{i=1}^{m} (\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(i)} - \hat{\boldsymbol{\mu}})^T$$

Note: Independence between X, Y implies $\mathrm{Cov}(X,Y) = 0$ but the inverse does not hold except for if X, Y are jointly normal.

1.2.2 Normal Distribution

The normal distribution in one dimension is characterised by the density

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\}$$

- Note: if $Z \sim \mathcal{N}(0,1)$, then $Y = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.
- We usually denote the density ϕ and the cdf by Φ.

We can extend the univariate normal to multiple dimensions. The multi-variate density is given by

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\mu)\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$

If $Z = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n), Z_i \sim \mathcal{N}(0,1)$ i.i.d. , then $Z \sim N(0, I_n)$. If $Y = \mu + AZ$, then

$$Y \sim \mathcal{N}(\mu, \Sigma) \text{ with } \Sigma = AA^T$$

The reason, why we often encounter the normal distribution is due to the central limit theorem.

Theorem 1.1 Classical Central Limit Theorem

If X_i are i.i.d. with $\mathbb{E}(X_i) = \mu, \text{Var}(X_i) = \sigma^2 < \infty$ then

$$\frac{1}{\sqrt{n}\sigma} \sum (X_i - \mu) \to \mathcal{N}(0,1) \text{ in distr as } n \to \infty$$

A more general statement is given by the Lindenberg CLT, which does not require identical but still independent distribution of our variables X_i:

Theorem 1.2 Lindenberg Central Limit Theorem

Suppose X_i are independently distributed random variables with $\mathbb{E}(X_i) = \mu_i, \text{Var}(X_i) = \sigma_i^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}\left[(X_i - \mu_i)^2 \cdot \mathbf{1}_{\{X_i | |X_i - \mu_i| > s s_n\}}\right] = 0,$$

then

$$\frac{1}{s_n} \sum_{i=1}^n (X_i - \mu_i) \to \mathcal{N}(0,1) \text{ (in distr as } n \to \infty$$

Note that if $X_1, \ldots, X_n \sim N(\mu, \sigma_X^2)$ i.i.d. then $\bar{X}_n \sim N\left(\mu, \frac{\sigma_X^2}{n}\right)$ exactly.

We next dicuss important properties of the normal distribution.

- If Y_i, Y_j are jointly normally distributed random variables, then

  $$\text{Cov}(Y_i, Y_j) = 0 \iff Y_i \perp Y_j$$

  I.e., zero covariance is equivalent to independence for normally distributed RVs[1]
- Linear transformations preserve normality. If $Y \sim \mathcal{N}(\mu, \Sigma)$ and A is a matrix of appropriate dimensions then

  $$AY = N(A\mu, A\Sigma A^T)$$

We also have the following more technical result:

Lemma 1.1 Independent linear transformations of a normal vector

Suppose $Y \sim \mathcal{N}(0, \sigma^2 I)$. Assume $A \in \mathbb{R}^{p\times n}, B \in \mathbb{R}^{q\times n}$ with $p + q \leq n$. Define $U = AY, V = BY$. Then

$$U \perp V \iff AB^T = 0$$

Proof. Write

$$\begin{bmatrix} U \\ V \end{bmatrix} = \begin{bmatrix} A \\ B \end{bmatrix} Y = CY$$

[1] Note that X, Y each being normal (but not jointly normal), is not sufficient. Counterexample:Let Z be standard normal and let $X = Z$ and $Y = Z$ if $|Z| \leq 1$, and $Y = -Z$ if $|Z| > 1$. Then X and Y are each normal, $Cov(X, Y) = 0$, but they are not independent.

then $CY \sim N(c_\mu, \Sigma)$. From normality, it follows that $U \perp V \iff \text{Cov}(U, V) = 0$. We have

$$\Sigma = \sigma^2 \begin{bmatrix} AA^T & AB^T \\ BA^T & BB^T \end{bmatrix}$$

Therefore, the covariance is zero if $AB^T = 0$ and since $\begin{bmatrix} U & V \end{bmatrix}^\top$ is multivariate normal, zero covariance implies independence. □

1.2.3 Distributions related to the normal

Chi-Squared.

- If $Z \sim \mathcal{N}(0, 1)$ then $Z^2 \sim \chi^2_1$
- If $Z_i \perp Z_j$ and each $Z_i \sim \mathcal{N}(0, 1)$ then $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$
- If $Y \sim \chi^2_n$ then $\mathbb{E}[Y] = n, \text{Var}(Y) = 2n$
- If $Y \sim \mathcal{N}(\mu, \Sigma)$ with $\dim(Y) = n$, then $(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2_n$

The following shows that if we have a $n \times n$ matrix with rank smaller than n (i.e., degenerate columns/rows), then only the actual rank 'matters' for the χ^2 distribution:

> **Lemma 1.2**
> Let $\mathbf{e} \sim \mathcal{N}(0, M)$, where M is an idempotent matrix of rank r. Then $\mathbf{e}^\top M \mathbf{e}$ has a central chi-square distribution with r degrees of freedom.

This follows from the fact that for a symmetric, idempotent matrix there exists an orthogonal matrix Γ such that $M = \Gamma D_\lambda \Gamma'$, where $D_\lambda = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$ with the number of ones being equal to the rank of the matrix M. Thus $\mathbf{e}' M \mathbf{e} = \mathbf{e}' \Gamma D_\lambda \Gamma' \mathbf{e} = \mathbf{z}' D_\lambda \mathbf{z} = \sum_{i=1}^r z_i^2$, where

$$\mathbf{z} = (z_1, \ldots, z_n)' = \Gamma' \mathbf{e} \sim N(0, \Gamma M \Gamma') = N(0, D_\lambda).$$

I.e., the vectors $\mathbf{z}$ are multivariate standard normal with the truncated identitiy matrix which is $D_\lambda = \text{diag}(1, \ldots, 1, 0, \ldots, 0)$. The matrix is not full rank, however the truncated matrix would be full-rank and is identical to the identitiy and thus identical to it's inverse. We can thsu apply our definition of the multi-variate χ^2 matrix from above.

F-distribution

- If $X \sim \chi^2_m, Y \sim \chi^2_n$ with $X \perp Y$, then

$$\frac{X/m}{Y/n} \sim F_{m,n}$$

- If $Y \sim F_{m,n}$ then $\mathbb{E}[Y] = n/(n-2)$

t-distribution

- If $Z \sim \mathcal{N}(0,1), X \sim \chi^2_k$ with $Z \perp X$ then

$$T = \frac{Z}{\sqrt{X/k}} \sim t_k$$

- The t-distribution is related to the F-distribution in a similar way as $\mathcal{N}$ is related to χ^2:

$$T \sim t_k \to T^2 \sim F_{1,k}$$

1.3 Statistics

1.3.1 MLE

In most cases, we want to use maximum likelihood estimators. MLEs have great asymptotic properties

- Consistent: $\hat{\theta} \to \theta$ in distribution as $n \to \infty$
- Asymptotically normal: $\sqrt{n}(\hat{\theta} - \theta) \to N\left(0, I^{-1}\right)$ in distribution as $n \to \infty$ where I denotes Fischer Information
- Efficient: MLE is asymptotically unbiased and has the smallest possible variance among reasonable estimators

1.3.2 Hypothesis Testing

z-Test

If we have normally distributed variables with known σ_x (or alternatively, if we can invoke a CLT to obtain asymptotic normality), then

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2_{\bar{x}}), \text{ with } \sigma_{\bar{x}} = \sigma_x/\sqrt{n}$$

We obtain our test and test result in five steps:

1. Specify null and alternalive hyptohesis:

$$H_0 : \mu = \mu_0, \; H_A : \mu \neq \mu_0$$

2. Test statistic:

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma_{\bar{X}_n}} = \frac{\sqrt{n}\left(\bar{X}_n - \mu_0\right)}{\sigma_X}$$

   Distribution of test statistcs under $H_0 : Z \sim \mathcal{N}(0,1)$

3. Choose level of signficance α (e.g. 0.05) and calculate rejection region ("tails" of null distribution):

$$K = \left(-\infty, -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right] \cup \left[\Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \infty\right) \text{ if } H_A : \mu \neq \mu_0$$

4. Finally, take samples, calculate test and reject or fail to reject

t-test

We usually do not know σ_x and need to estimate it from data. We proceed as before but modify the test statistic and rejection region based on the new distr. of the test statistic:

$$\mathrm{T} = \frac{\bar{X}_n - \mu_0}{\hat{\sigma}_{\bar{X}_n}} = \frac{\sqrt{n}\left(\bar{X}_n - \mu_0\right)}{\hat{\sigma}_X}$$

where $\hat{\sigma}_X^2 = \frac{1}{n-1}\Sigma\left(X_i - \bar{X}_n\right)^2$.

- Distribution of test statistcs under $H_0 : T \sim t_{n-1}$.
- Rejection region for test statistics:

$$K = \left(-\infty, -qt\left(n-1; 1-\frac{\alpha}{2}\right)\right] \cup \left[qt\left(n-1; 1-\frac{\alpha}{2}\right), \infty\right) \text{ if } H_A : \mu \neq \mu_0$$

Likelihood Ratio Test

Due to the Neyman-Pearson lemma, we know that likelihood ratio tests have largest power for deciding between two concrete single hypothesis. Asymptotically, LR tests are χ^2 distributed.

Definition 1.13 Likelihood ratio test

Define λ by

$$\lambda = \frac{L\left(\hat{\Omega}_0\right)}{L(\hat{\Omega})} = \frac{\sup_{\Theta\in\Omega_0} L(\Theta)}{\sup_{\Theta\in\Omega} L(\Theta)}.$$

A likelihood ratio test of $H_0 : \Theta \in \Omega_0$ versus $H_a : \Theta \in \Omega_a$ employs λ as a test statistic, and the rejection region is determined by $\lambda \leq k$ where k is chosen to achieve the desired significance level α.

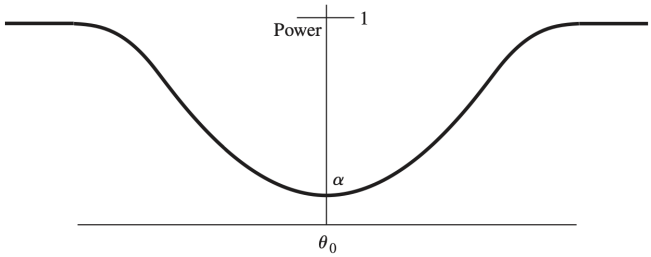
Where we use the following notation: null hypothesis specifies that Θ lies in a particular set of possible values Ω_0 and the alternative hypothesis specifies that Θ lies in another set of possible values Ω_a, which does not overlap Ω_0. Let $L\left(\hat{\Omega}_0\right) = \sup_{\Theta\in\Omega_0} L(\Theta)$ and $L(\hat{\Omega}) = \sup_{\Theta\in\Omega} L(\Theta)$.

Type I, Type II error and power

Definition 1.14

A **type I error** is made if H_0 is rejected when H_0 is true. The probability of a type I error is denoted by α. The value of α is called the **level** of the test.
A **type II error** is made if H_0 is accepted when H_a is true. The probability of a type II error is denoted by β.

Figure 1: A typical power curve



I.e.,

$$\begin{aligned}\beta &= P(\text{ type II error }) = P(\text{ accepting } H_0 \text{ when } H_a \text{ is true }) \\ &= P(\text{ value of the test statistic is not in RR when } H_a \text{ is true }).\end{aligned}$$

Clearly, α, β are inversely related as the higher α, the smaller the rejection region and thus the higher the possibillity of not rejecting the null even thoguh H_0 is wrong.

The trade-off between the two errors is closely related to a concept called power.

Definition 1.15 Power of a test

Suppose that W is the test statistic and RR is the rejection region for a test of a hypothesis involving the value of a parameter θ. Then the power of the test, denoted by power (θ), is the probability that the test will lead to rejection of H_0 when the actual parameter value is θ. That is, power$(\theta) = P(W$ in RR when the parameter value is $\theta)$.

If we write a decision function for our test such that $d(W) = 1$ if $W \in RR$ and 0 otherwise, then we can write

$$\text{power}(\theta) = \mathbb{E}_\theta(d(W)) \tag{1.2}$$

If θ_a is some value in H_a, i.e. a value not equal to the null hypothesis, then

$$\text{power}\,(\theta_a) = 1 - \beta\,(\theta_a)$$

P-value

Definition 1.16 P-Value

Consider an observed test-statistic t from unknown distribution T. Then the p-value p is what the prior probability would be of observing a test-statistic value at least as "extreme" as t if null hypothesis H_0 were true.

Note that under the null hypothesis, the p-value is uniformly distributed over $[0, 1]$. For simplicity, consider a one-sided t-test. Then $p = \Pr(T \geq t) = 1 - F(t)$ where F is the cdf of the distribution of the test statistic under H_0. The *Probability Integral Transform* tells us

that the random variable $Z_t := F(t)$ is uniformly distributed over $[0, 1]$ for any cumulative distribution function.

Confidence Intervals

CI show a set of "compatible" parameter values. A $(1-\alpha)$-confidence interval contain all parameter values for which a test at significance level α would fail to reject.

We can often solve for CIs analytically: Find all values of μ, so that the test does *not* reject, i.e.:

$$\left| \frac{\sqrt{n}\,(\bar{x}_n - \mu)}{\partial_X} \right| \leq qt\left(n-1; 1-\frac{\alpha}{2}\right) \rightarrow \text{ solve for } \mu$$

Thus, $(1-\alpha)$-confidence interval is:

$$\left(\bar{X}_{\mathrm{n}} - \mathrm{qt}\left(\mathrm{n}-1; 1-\frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}; \bar{X}_{\mathrm{n}} + \mathrm{qt}\left(\mathrm{n}-1; 1-\frac{\alpha}{2}\right) \cdot \frac{\hat{\sigma}_X}{\sqrt{n}}\right)$$

- Rule of thumb for 95%-CI: $\left(\bar{X}_{\mathrm{n}} - 2 \cdot \frac{\partial_X}{\sqrt{n}}; \bar{X}_{\mathrm{n}} + 2 \cdot \frac{\partial_X}{\sqrt{n}}\right)$
- A $(1-\alpha)$-confidence interval covers the true value with probability $(1-\alpha)$.

Multiple Testing

Suppose we consider performing 100 T-tests at a significance level $\alpha = 0.05$. Even if there is no effect, we expect to find about 5 significant results. One way of correcting for this is considering a conservative test level such that the probability of at least one type-1 error is $\leq \alpha$. This is called a **Familywise Error Rate**.

One simple and conservative approach is the **Bonferroni correction**: test at level α/m if m tests are performed.

1.3.3 Measuring association

Pearon correlation is a standardized measure of the covariance and measures linear association:

$$\rho_{X,Y} = \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

For inference, we may use the *Fisher z-transform* which is given by

$$Z := \tanh^{-1}(\widehat{\rho}) = \frac{1}{2}\log\left(\frac{1+\widehat{\rho}}{1-\widehat{\rho}}\right) \tag{1.3}$$

Proposition 1.13

If $\mathbf{X}, \mathbf{Y}$ follows a multivariate normal distribution, then Z as defined in (1.3) is ap-

proximately

$$Z \sim \mathcal{N}\left(\tanh^{-1}(\rho), \frac{1}{n-3}\right)$$

A 95%-CI for ρ is thus given by

$$\tanh\left(z \pm 1.96 \cdot \sqrt{\frac{1}{n-3}}\right)$$

If we want to measure nonlinear associations as well, one possiblility is **Spearman correlation** which measures montone relationships. r_S can be obtained by computing ranks and then applying Pearson correlation, or using

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} D_i^2}{n\left(n^2-1\right)}, \quad D_i := \operatorname{rank}\left(x_i\right) - \operatorname{rank}\left(y_i\right)$$

Simple Linear Regression

Consider estimating a linear relation between pairs of variables (x_i, Y_i): I.e., we want to find a line of the form $Y_i = \beta_0 + \beta_1 x_i$ that is as close as possible to y_i. We can fit our model by minimizing the sum of squared residuals, i.e.,

$$\widehat{\beta_0}, \widehat{\beta_1} = \arg\min \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_i\right)\right)^2$$

The solution is given by

$$\hat{\beta}_1 = \hat{\rho}_{XY} \cdot \frac{\widehat{\sigma}_Y}{\hat{\sigma}_X}, \quad \hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n$$

2 Estimating the Linear Regression Model

Assume X are fixed and $\text{rank}(X) = p$ where p is the number of predictors (including the intercept). Then

$$Y_i = \beta_1 + \beta_2 X_i^{(2)} + \ldots + + \beta_p X_i^{(p)} + \epsilon_i$$

We may write our regression equation in vector or matrix form as

$$\begin{aligned} Y_i &= X_i^\top \beta + \varepsilon_i, \\ Y &= X\beta + \varepsilon \end{aligned}$$

with $\mathbb{E}(\varepsilon_i) = 0, \text{Var}(\varepsilon_i) < \infty$.

One way to estimate $\hat{\beta}$ from data is by minimizing the residual sum of squares:

$$\min \sum_{i=1}^{n} \left(Y_i - X_i^\top \beta\right)^2$$

This problem is equivalent to MLE with $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{1})$.

In matrix notation, we may write the objective function as

$$(Y - X\beta)^\top (Y - X\beta)$$

Taking gradient and setting equal to zero yields the normal equations:

$$X^T(Y - X\hat{\beta}) = 0 \to X^T X\hat{\beta} = X^T Y$$

The closed-form solution is then given by

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \tag{2.1}$$

To see equivalence between OLS and MLE under $\epsilon \sim \mathcal{N}$, note that the log likelihood of our data is given by

$$\log L = -n \cdot \log(\sigma\sqrt{2\pi}) - \frac{1}{2} \frac{\left(\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1,i} - \cdots - \beta_{p-1} x_{p-1,i})^2\right)}{\sigma^2}$$

clearly this quantitiy is maximized when the sum of squared residuals is minimized. One can also see from the quadratic function that we have a convex (or concave) optimization problem.

2.1 Variable transformations

Linear regression only requries linearity in parameters. We can for instance apply complex transformations to our X columns and our target variable Y.

We often encounter the following variable transformations in our regression models:

1. **Exponential type**:

$$y = \exp(\beta_0 + \beta_1 \cdot x) \cdot \tilde{\varepsilon} \to \tilde{y} = \beta_0 + \beta_1 \cdot x + \varepsilon \text{ with } \tilde{y} = \log(y)$$

2. **Power type**:

$$y = \beta_0 \cdot x^{\beta_1} \cdot \tilde{\varepsilon} \to \tilde{y} = \widetilde{\beta_0} + \beta_1 \cdot \tilde{x} + \varepsilon$$

with $\tilde{y} = \log(y), \tilde{x} = \log(x), \widetilde{\beta_0} = \log(\beta_0)$

In both case, we initially have a multiplicative error which we transform into an additive error.

Note, in general $g^{-1}(\mathbb{E}(g(Y))) \neq \mathbb{E}(Y)$. In particular, $\exp(E(\log(Y))) \neq E(Y)$. We therefore need to be careful when interpreting the results of our transformed equation back on the original scale.

However, if our distribution on the transformed scale is symmetric and since logarithms preserve ordering, $\mathbb{E}(\log(Y)) = \text{median}(\log(Y)) = \log(\text{median}(Y))$. Thus, $\exp(\mathbb{E}(\log(Y))) = \text{median}(Y)$. The effect interpretation of our coefficient is then

$$\frac{\exp\left(E\left(\log(Y) \mid x_i + 1\right)\right)}{\exp\left(E\left(\log(Y) \mid x_i\right)\right)} = \exp\left(\tilde{\beta}_1\right) = \frac{\text{Median}\left(Y \mid x_i + 1\right)}{\text{Median}\left(Y \mid x_i\right)}$$

I.e., "It is estimated that the median of Y given $x_i + 1$ is $\exp\left(\tilde{\beta}_1\right)$ times as large as the median of Y given x_i."

2.2 Linear Regression as Orthogonal Projection

If we interpret our data in X 'column-wise', our columns are vectors spanning a (p-dimensional) subspace of $\mathbb{R}^n$.[2] I.e., the vector Y of observations is a single point in the n-dimensional space $\mathbb{R}^n$. If we vary the value of the parameter β, the product $X\beta$ describes a p-dimensional hyperplane through the origin.

We can obtain our OLS estimates by trying to find the element $X\beta$ on the hyperplane which lies closest to the point Y. I.e.,

$$\text{Choose } \beta \text{ to minimize } L_2\text{-norm: } |Y - X\beta|_2{}^2$$

Since we obtain the orthogonal projection of Y onto hyperplane spanned by columns of X, the residual vector $\hat{\varepsilon} = y - X\hat{\beta}$ must be orthogonal to any vector in the hyperplane, see Figure 2 for an illustration. With this, we recover the normal equations $X^T(Y - X\hat{\beta}) = 0$. Solving for $\hat{\beta}$ yields

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

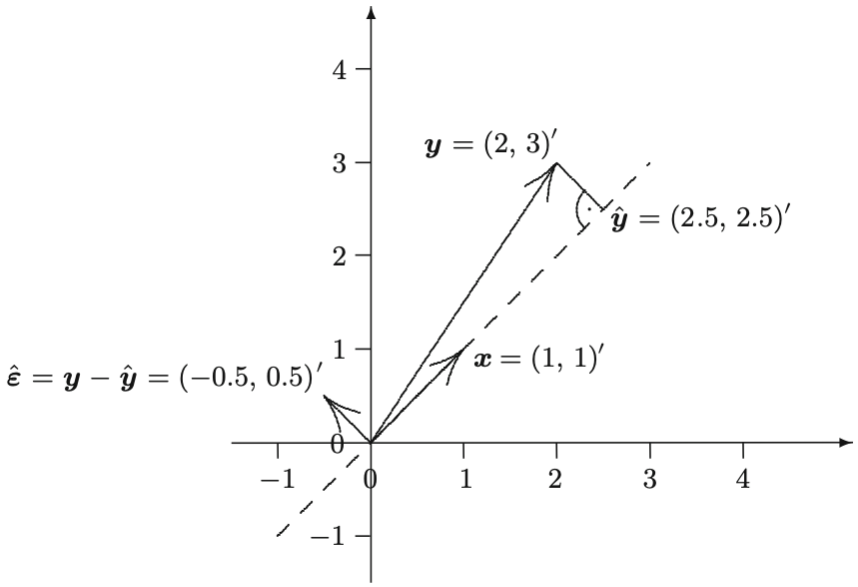
Our projection of Y onto this subspace is therefore

$$\hat{Y} = X\hat{\beta} = \underbrace{X\left(X^T X\right)^{-1} X^T}_{:=H} Y = HY$$

Note that H has the exact form of the orthogonal projection matrix in (1.1).

[2] Alternatively, we can interpret our data and regression 'row-wise' in which our features and target are each a different basis of $\mathbb{R}^{p+1}$. We then want to fit a hyperplane that is as close as possible to all of our points.

Figure 2: Graphical Intuition for Linear Regression as Projection



If we want to obtain our residuals, we can do so by using the matrix $M = 1 - H$, since

$$MY = (1 - H)Y = Y - \hat{Y} = \hat{\varepsilon}$$

This is also a projection matrix. We also know (assuming $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$)

$$\hat{Y} \perp \hat{\varepsilon} \text{ since } H \cdot M = 0 \tag{2.2}$$

which follows from Lemma 1.1. Also,

$$X^T \hat{\varepsilon} = 0. \tag{2.3}$$

Intuitively, we can see (2.3) by noting that $\hat{\varepsilon}$ is orthogonal to the column space of X. Formally, we can show this as

$$\begin{aligned} X^T \hat{\varepsilon} &= X^T (y - \hat{y}) \\ &= X^T (1 - H) y \\ &= X^T (1 - X(X^T X)^{-1} X^T) y \\ &= X^T y - (X^T X)(X^T X)^{-1} X^T y \\ &= 0 \end{aligned}$$

This result leads to the following useful consequences:

- $\hat{\beta} \perp \hat{\varepsilon}$
- $\hat{\varepsilon} = MY = M(X\beta + \varepsilon) = M\varepsilon$

- If our model includes an intercept, one column of X is $E = (1, 1, \ldots, 1)^T$. Thus

$$E \cdot \hat{\varepsilon} = \sum \hat{\varepsilon}_i = 0$$

- $\frac{1}{n} \sum_{i=1}^{n} \hat{y}_i = \bar{y}$.
- The regression hyperplane runs through the average of the data, i.e.,

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k.$$

For the first of the above results, note

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y = \left(X^T X\right)^{-1} X^T (\hat{Y} + \hat{\varepsilon}) = \left(X^T X\right)^{-1} X^T \hat{Y}$$

so $\hat{\beta}$ is only a function of X and $\hat{Y}$ and $\hat{Y} \perp \hat{\varepsilon}$.

Our projection interpretation allows us to perform the following Pythagoras decomposition:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}$$

- Called *coefficient of determination*
- Measure of goodness of fit
- Note that $|\hat{Y} - \bar{Y}|$ lies in the column space of X and $|Y - \hat{Y}|$ lies in the orthogonal complement. Thus these two vectors are two sides to a right-angled triangle so

$$TSS = ESS + RSS$$

  where $TSS = |Y - \bar{Y}|^2$
- One can show that $R^2 = \left[cor(Y, \hat{Y})\right]^2$.

2.3 Simple vs Multiple Regression and Partial Correlation

Unless predictors are orthogonal, many simple regressions will yield different coefficient from multiple regression. Interpretation of multiple regression: Effect of x_k on Y when keeping x_{-k} fixed.

Special case: Orthogonal design

- Orthogonal predictors: $x_{\cdot,j}^{\top} x_{\cdot,k} = 0$ for all $j \neq k$
- $X^T X = \text{diag}\left(\sum_{i=1}^{n} x_{i,1}^2, \ldots, \sum_{i=1}^{n} x_{i,p}^2\right)$ Thus:

$$\hat{\beta}_j = \left(\left(X^T X\right)^{-1} X^T Y\right)_j = \frac{\sum_{i=1}^{n} x_{i,j} y_i}{\sum_{i=1}^{n} x_{i,j}^2} = \frac{\widehat{\sigma}_{x_j, y}}{\hat{\sigma}_{x_j}^2}$$

  which is the same result as simple regressions

We can therefore view multiple regression as measuring *partial correlation*. An informal definiton of partial correlation is:

Partial correlation $\rho_{XY|Z}$: strength and direction of the linear dependence between X and Y after accounting for the linear dependence of X and Y on Z

One can estimate partial correlation in the following ways:

1. A recursive formula
2. Let r_x be residuals when regressing x on z, r_y residuals when regressing y on z, then $\rho_{XY|Z} = \text{cor}\left(r_x, r_y\right)$
3. Via precision matrix, $\rho_{YX^j|X^{-j}} = -\frac{K_{p+1,j}}{\sqrt{K_{p+1,p+1}K_{j,j}}}$ where $K = \Sigma^{-1}$ and $\Sigma = \text{Cov}\left((X_{*,1}, \ldots, X_{*,p}, Y)^T\right)$

In multiple linear regression, we can write our estimated coefficients using partial correlation as

$$\beta_j = -\frac{K_{p+1,j}}{K_{p+1,p+1}}; \quad \rho_{YX^j|X^{-j}} = \beta_j \frac{\sqrt{K_{p+1,p+1}}}{\sqrt{K_{j,j}}} \tag{2.4}$$

2.4 Computational considerations for estimating linear regression

Instead of directly using the closed-form solution in (2.1), we can estimate the OLS coefficients in different ways. From (2.4) we can see that we can estimate them using partial correlations. The numerically prefered way is using a QR decomposition of X. I.e., let $X = QR$ where Q is $n \times p$ with orthonormal columns and R is $p \times p$ upper triangle. From the normal equations, we have

$$\begin{aligned}
& X^T X\hat{\beta} = X^T y \\
& (QR)^T (QR)\hat{\beta} = (QR)^T y \\
& R^T Q^T QR\hat{\beta} = R^T Q^T y \\
& R^T R\hat{\beta} = R^T Q^T y \\
& R\hat{\beta} = Q^T y \\
\implies & \hat{\beta} = R^{-1} Q^T y
\end{aligned}$$

2.5 Categorical Variables and ANOVA

We can include categorical variables as predictors by using dummy encoding (i.e., factors).

- E.g., we can allow for a different intercept between two groups, where x_2 is a binary variable:
$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}(x_2 = 1) + \epsilon_i$$
- One reference level and $k - 1$ binary dummies if we have k categories.

- Contrasts to derive the intercept in differnt groups. Let $\ell = \begin{pmatrix} 1 & 0 & 1 \end{pmatrix}^\top$. Then $\ell^\top \beta$ is the intercept for the group with $x_2 = 1$.
- Interactions to allow for different slopes between groups. E.g.,

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}(x_2 = 1) + \beta_3 x_1 \cdot \mathbf{1}(x_2 = 1) + \epsilon_i$$

Analysis of Variance (ANOVA) is a special case of linear regression which is closely related to categorical variables. We illustrate the idea behind ANOVA using a simple example.

Let $g = 4$ be the number of groups (or 'treatments') and p denote the number of observations per group which we assume to be the same in each group. Our null hypothesis is

$$H_0 : \ \mu_1 = \mu_2 = \mu_3 = \mu_4$$

To test this, we calculate:

- **Variation between groups**:

$$SS_B = p \cdot \sum_{i=1}^{g} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot} \right)^2$$

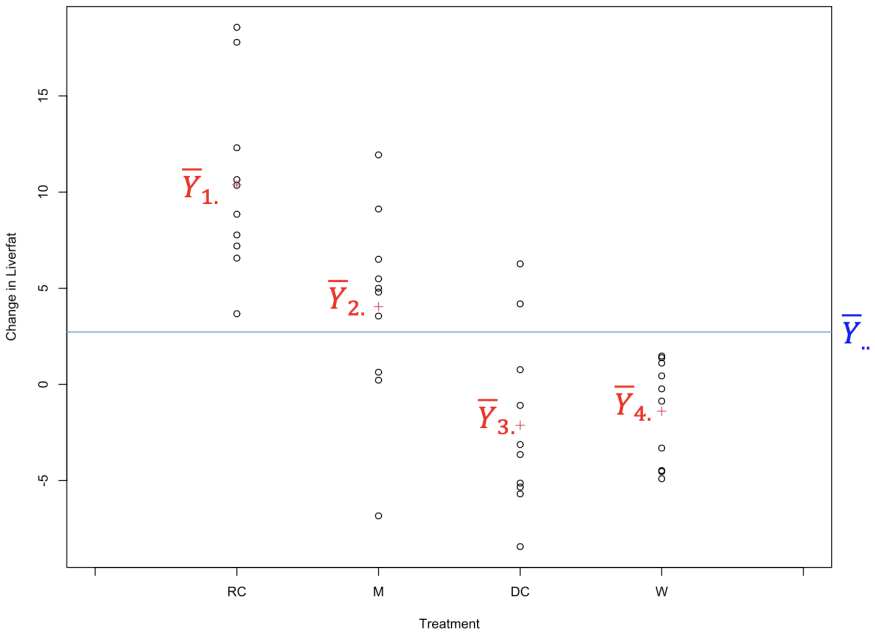
  where $\bar{Y}_{\cdot\cdot}$ is the overall mean of all of observations.

- **Variation within groups**:

$$SS_W = \sum_{i=1}^{g} \sum_{j=1}^{p} \left(Y_{ij} - \bar{Y}_{i\cdot} \right)^2$$

Our test statistic is then approximately $\dfrac{SS_B}{SS_w}$ which follows an F-distribution under suitable assumptions. We reject our null if the variation between group averages is much larger than within group variation. The setup is illustrated in Figure 3.

Figure 3: Example ANOVA setup



3 Properties and Efficiency of OLS Estimates

Note that our coefficients depend on Y and, as Y is random, are random variables themselves. We must therefore think about precision and distribution of estimated parameters. We assume the *Gauss-Markov Conditions*:

$$Y = X\beta + \varepsilon$$
$$E(\varepsilon) = 0, \text{Cov}(\varepsilon) = \sigma^2 I$$

We obtain the following moments:

- $\hat{\beta}$:
  - $\mathbb{E}\left[\hat{\beta}\right] = \beta$
  - $\text{Cov}\left(\hat{\beta}\right) = \sigma^2 (X^T X)^{-1}$
- $\hat{Y}$:
  - $\mathbb{E}\left[\hat{Y}\right] = \mathbb{E}\left[Y\right] = X\beta$
  - $\text{Cov}\left(\hat{Y}\right) = \sigma^2 H$
- $\hat{\varepsilon}$:
  - $\mathbb{E}\left[\hat{\varepsilon}\right] = 0$
  - $\text{Cov}\left(\hat{\varepsilon}\right) = \sigma^2 M$ which implies $\text{Var}(\hat{\epsilon}_i) = \sigma^2(1 - H_{ii})$
  - $\text{Cov}\left(\hat{\varepsilon}, \hat{Y}\right) = 0$

For the variance of a single element of our coefficient vector, one can obtain the following formula

$$\text{Var}\left(\hat{\beta}_j\right) = \frac{\sigma^2}{\left(1 - R_j^2\right) \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2},$$

where R_j^2 is the coefficient of determination from regressing x_j as the response on all other variables in X. We note:

- The smaller the model variance σ^2, the smaller the variance of $\hat{\beta}_j$ and thus the more accurate the estimation.
- The smaller the linear dependence between x_j and the other explanatory variables (measured through R_j^2), the smaller is the variance of $\hat{\beta}_j$. The variance, $\text{Var}\left(\hat{\beta}_j\right)$, is minimized for $R_j^2 = 0$, i.e., when the covariates are uncorrelated.
- The larger the variability of covariate x_j around its average, the smaller is the variance of $\hat{\beta}_j$.

Further, we can show that $1/(n-p)\sum_{i=1}^{n}\hat{\epsilon}_i^2$ is an unbiased estimate of σ^2. To this end, note that

$$\begin{aligned}
\mathbb{E}\left[\hat{\sigma}^2\right] &= \frac{1}{n-p}\sum_{i=1}^{n}\mathbb{E}\left[\hat{\epsilon}_i^2\right] = \frac{1}{n-p}\sum_{i=1}^{n}\mathrm{Var}\left[\hat{\epsilon}_i^2\right] \\
&= \frac{1}{n-p}\sum_{i=1}^{n}\sigma^2(1-H_{ii}) = \frac{1}{n-p}\sigma^2\left(n-\mathrm{tr}(H)\right) \\
&= \frac{1}{n-p}\sigma^2\left(n-\mathrm{tr}(X(X^TX)^{-1}X)\right) \\
&= \frac{1}{n-p}\sigma^2\left(n-\mathrm{tr}(\mathbf{1}_{p\times p})\right) = \sigma^2
\end{aligned}$$

Note here also the following insight into the relation between degrees of freedom and projection matrices:

- We have n observations and estimate p parameters to obtain residuals. Thus residual df are $n-p$
- The trace of a projection matrix measures the rank of the subspace onto which it projects a vector. I.e., how many independent dimensions there are in a subspace. We have $\mathrm{tr}(M) = tr(\mathbf{1}-H) = n-p$.

3.1 Gauss-Markov Theorem

Given data Y, a parameter estimate in general is some function $\hat{\beta} = f(Y)$. We want to understand goodness properties of OLS and how (or, whether) there are ways to improve upon OLS.

Note that OLS has the following properties

1. $\hat{\beta} = (X^TX)^{-1}X^Ty = Ay$ is *linear*
2. $\mathbb{E}(\hat{\beta}) = \beta$, i.e., *unbiased*

We will show that, depending on more or less restrictive assumptions, OLS can be

- Optimal among all *unbiased estimators*, and
- Optimal among all *linear and unbiased* estimators.

In both cases we may obtain an estimator with lower MSE by allowing for small bias if we can achieve a large reduction of variance simultaneously, see equation (3.4).

These results and the underlying assumptions are summarised in the following two theorems.

Theorem 3.1 Gauss-Markov – Version 1

Let $Y = X\beta + \varepsilon, E(\varepsilon) = 0, \mathrm{Cov}(\varepsilon) = \sigma^2 I, rnk(X) = p$. Furthermore, let $\ell \in \mathbb{R}^p$, and $\hat{\beta}$ the OLS estimator.

Then, for all $c \in \mathbb{R}^n$ such that $E\left(c^T y\right) = \ell^T \beta$, we have:

$$\text{Var}\left(\ell^T \hat{\beta}\right) \leq \text{Var}\left(c^T y\right).$$

Thus, the OLS estimator $\ell^T \hat{\beta}$ has minimal variance among all linear unbiased estimators of $\ell^T \beta$.

Proof. Let $y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}, \beta \in \mathbb{R}^p$. The OLS estimator is given by

$$\begin{aligned} \hat{\beta} &= \left(X^T X\right)^{-1} X^T y \\ \implies \ell^T \hat{\beta} &= \ell^T \left(X^T X\right)^{-1} X^T y \end{aligned}$$

We know that $\mathbb{E}\left[\ell^T \hat{\beta}\right] = \ell^T \beta$. Now let $c^T y$ be any other linear unbiased estimator of $\ell^T \beta$. Since this estimator is unbiased,

$$\ell^\top \beta = \mathbb{E}\left[c^T y\right] = c^T X \beta$$

so that we must have $\ell^T = c^T X$.

We now consider variances.

$$\begin{aligned} \text{Var}\left(c^T y\right) &= \sigma^2 c^T c \\ \text{Var}\left(\ell^T \beta\right) &= \sigma^2 \ell^T \left(X^T X\right)^{-1} \ell \\ &= \sigma^2 c^T X \left(X^T X\right)^{-1} X^T c \end{aligned}$$

Subtracting both variances, we obtain

$$\begin{aligned} &\text{Var}\left(c^T y\right) - \text{Var}\left(\ell^T \beta\right) \\ &= \sigma^2 c^T \left(1 - X^T \left(X^T X\right)^{-1} X^T\right) c \\ &= \sigma^2 c^T M c \geq 0 \end{aligned}$$

since M is positive semi-definite.[3] □

If we are further willing to assume normality of the error term, we arrive at a stronger result.

Theorem 3.2 Gauss-Markov – Version 2

Let furthermore ε be normally distributed. Then $\ell^T \hat{\beta}$ has minimal variance among

[3] Since M is a orthogonal projection, $M^T = M = M^2$. Let z be any vector of appropriate dimensions, then

$$z^T M z = z^T M^T M z = (Mz)^T Mz = |Mz|^2 \geq 0$$

all unbiased estimators of $\ell^T\beta$.

To proof the above theorem, we will use the Cramer-Rao bound result:

Let $(f_\eta(\mathbf{Y}))$ be a parametric family of strictly positive densities in $\mathbb{R}^n$. Let η be a variable parameter with values in an open subset of $\mathbb{R}^k$, and let $f_\eta(\mathbf{Y})$ be differentiable wrt η. Our parameter of interest is $g(\eta)$, where g is an arbitrary real-valued function of η. Then we obtain the following result:

Theorem 3.3 Cramér-Rao for unbiased estimators

If $T(\mathbf{Y})$ is an arbitrary unbiased estimate of $g(\eta)$, i.e.

$$\mathbb{E}_\eta[T(\mathbf{Y})] = g(\eta) \quad \forall \eta,$$

then g is differentiable and

$$\mathrm{Var}_\eta(T(\mathbf{Y})) \geq \frac{\partial g^T}{\partial \eta} I(\eta)^{-1} \frac{\partial g}{\partial \eta} \tag{3.1}$$

where $I(\eta)$ denotes the Fisher information matrix:

$$I(\eta) = \mathbb{E}_\eta \left[\frac{\partial \log f_\eta(\mathbf{Y})}{\partial \eta} \frac{\partial \log f_\eta(\mathbf{Y})^T}{\partial \eta} \right].$$

Proof. We know that

$$\mathrm{Var}\left(\ell^T\beta\right) = \ell^T\sigma^2(X^TX)^{-1}\ell \tag{3.2}$$

We now derive the Cramer-Rao lower bound. We have $\eta = (\sigma^2, \beta^T)^T$ and $g(\eta) = \ell^T\beta$. Then the Fisher information matrix is given by

$$I(\eta) = \mathbb{E}_\eta \left[\nabla_\eta \log L \cdot \nabla_\eta \log L^T \right]$$

where we use $\log L$ to denote the log likelihood function.

Our log likelihood function is

$$\log L = \sum_{i=1}^N -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{\left(y_i - x_i^T\beta\right)^2}{2\sigma^2}$$

Taking derivatives,

$$\nabla_\eta \log L = \begin{bmatrix} \sum_{i=1}^N \left(-\frac{1}{2\sigma^2} + \frac{\left(y_i - x_i^T\beta\right)^2}{2\sigma^4} \right) \\ \sum_{i=1}^N \left(\frac{\left(y_i - x_i^T\beta\right)x_i}{\sigma^2} \right) \end{bmatrix} \tag{3.3}$$

Since observations are independent, the Fisher information matrix is the sum of the individual Fisher information matrices for single observations. We compute,

$$I_{11}^{(i)} = \mathbb{E}\left[\left(-\frac{1}{2\sigma^2} + \frac{\left(y_i - x_i^T\beta\right)^2}{2\sigma^4}\right)^2\right] = \frac{1}{2\sigma^4}$$

$$I_{jk}^{(i)} = \mathbb{E}\left[\frac{\left(y_i - x_i^T\beta\right)^2 x_j x_k}{\sigma^4}\right] = \frac{x_j x_k}{\sigma^2}$$

$$I_{1j}^{(i)} = I_{j1}^{(i)} = \mathbb{E}_\eta\left[\left(-\frac{1}{2\sigma^2} + \frac{\left(y - x^T\beta\right)^2}{2\sigma^4}\right) \cdot \frac{\left(y - x^T\beta\right) \cdot x_j}{\sigma^2}\right] = 0$$

For the first result, we use $\mathbb{E}\left[\left(y_i - x_i^T\beta\right)^4\right] = 3\sigma^4$ for the normal distribution. In the last result, after extending and simplyfing, we again use this result and the unbiasedness of our estimate.

Summing over all observations, we arrive at

$$I(\eta) = \begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{1}{\sigma^2}X^TX \end{bmatrix}$$

Taking the inverse of this block matrix, we get

$$I(\eta)^{-1} = \begin{bmatrix} \frac{2}{n}\sigma^4 & 0 \\ 0 & \sigma^2(X^TX)^{-1} \end{bmatrix}$$

We compute the gradient vectors:

$$\frac{\partial g}{\partial \eta} = \begin{bmatrix} \partial \ell^T\beta/\partial\sigma^2 \\ \partial \ell^T\beta/\partial\beta_1 \\ \vdots \\ \partial \ell^T\beta/\partial\beta_p \end{bmatrix} = \begin{bmatrix} 0 \\ \ell_1 \\ \vdots \\ \ell_p \end{bmatrix} = \begin{bmatrix} 0 \\ \ell \end{bmatrix}$$

Finally, multiplying out, we obtain

$$\frac{\partial g^T}{\partial \eta} I(\eta)^{-1} \frac{\partial g}{\partial \eta} = \ell^T\sigma^2\left(X^TX\right)^{-1}\ell$$

so that our OLS estimator's variance in (3.2) achieves the Cramer-Rao lower bound in (3.1). □

Note that there may be biased estimators which yield a lower mean squared error. In particular, the MSE can be decomposed as

$$\mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \left(\mathbb{E}(\hat{\theta}) - \theta\right)^2 + \mathrm{Var}(\hat{\theta})\Big) \tag{3.4}$$

When may our OLS estimates perform badly?

- If our normality assumption holds, one may obtain slightly biased estimators with far lower variance
- If normality assumption is violated, OLS and MLE no longer coincide so we no longer have asymptotic efficiency. Non-linear estimates may outperform OLS.

3.2 Distribution of $\hat{\beta}$

Normal errors

Under normal errors, MLE coincides with OLS, so we should expect asymptotic nromallity of our estimated coefficients. We can, however, show that this already holds in finite samples.

Proposition 3.1

Assume that $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$. Then

1. $\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)$
2. $\widehat{\mathrm{Y}} \sim \mathcal{N}_n\left(X\beta, \sigma^2 H\right), \quad \widehat{\varepsilon} \sim \mathcal{N}_n\left(\mathbf{0}, \sigma^2 M\right)$
3. $\widehat{\mathrm{Y}}$ and $\widehat{\varepsilon}$ are independent
4. $\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2} \sim \chi^2_{n-p}$
5. $\widehat{\sigma}^2$ is independent of $\widehat{\beta} = \left(X^T X\right)^{-1} X^T Y$

Note that result 2) implies that even though ε are i.i.d., the covariance structure of $\hat{\varepsilon}$ can be highly dependent, including heteroskedasticity.

Proof. 1.) Since $Y \sim N(X\beta, \sigma^2 \mathbb{I})$ and $\hat{\beta} = (X^T X)^{-1} X^T y$, this implies that $\hat{\beta} \sim N(.,.)$. Note that

$$\mathbb{E}\left[\hat{\beta}\right] = (X^T X)^{-1} X^T \mathbb{E}\left[y\right] = (X^T X)^{-1} X^T (X\beta + \mathbb{E}\left[\varepsilon\right]) = \beta$$

and

$$\mathrm{Var}\left(\hat{\beta}\right) = \left[(X^T X)^{-1} X^T\right] \mathrm{Var}\left(y\right) \left[(X^T X)^{-1} X^T\right]^T = \sigma^2 (X^T X)^{-1}$$

2.)

$\hat{Y} = HY$ where $H = X(X^T X)^{-1} X^T$ with $\mathrm{rank}(H) = p$. Thus $\hat{Y} \sim N(.,.)$. We have

$$\mathbb{E}\left[\hat{Y}\right] = H\mathbb{E}\left[Y\right] = HX\beta = X\beta = \mathbb{E}\left[Y\right]$$

This follows since H is a projection onto the column space of X and $X\beta$ evidently already lies in the col. space of X. Also,

$$\mathrm{Var}\left(\hat{Y}\right) = H \,\mathrm{Var}\left(Y\right) H^T = \sigma^2 H H^T = \sigma^2 H$$

$\hat{\varepsilon} = MY$ where $M = \mathbb{I} - H$.

$$\begin{aligned} \mathbb{E}\left[\hat{\varepsilon}\right] &= M\mathbb{E}\left[y\right] = MX\beta = (\mathbb{I} - H)X\beta = 0 \\ \mathrm{Var}\left(\hat{\varepsilon}\right) &= \mathrm{Var}\left(MY\right) = \sigma^2 M \end{aligned}$$

3.)

$\hat{Y} = HY, \hat{\varepsilon} = MY$. Note that

$$HM = 0$$

Thus, by Lemma 1.1, we have $\hat{Y} \perp \hat{\varepsilon}$

4.)

Let $e_i = \hat{\varepsilon}_i / \sigma$. Then

$$\frac{\sum \hat{\varepsilon}_i}{\sigma^2} = e^T e = e^T M e$$

We can multiply by the matrix M since e lies in the orthogonal complement of the column space of X and is therefore unaffected by multiplying with the orthogonal projection matrix M.

We know that $e_i \sim N, \mathbb{E}[e] = 0, \text{Var}(e) = \frac{1}{\sigma^2}\text{Var}(\hat{\epsilon}_i) = M$. Therefore, $e \sim \mathcal{N}(0, M)$. We also know that M is idempotent and

$$\text{rank}(M) = \text{rank}(\mathbb{I} - H) = n - p$$

Then, using Lemma 1.2, $e^T M e \sim \chi^2_{n-p}$.

5.)

Since $X^T \hat{e} = 0$.

$$\hat{\beta} = (X^T X)^{-1} X^T (\hat{Y} + \hat{\varepsilon}) = (X^T X)^{-1} X^T \hat{Y}.$$

Thus, $\hat{\sigma}$ is a function of $\hat{\varepsilon}$ and $\hat{\beta}$ is a function of $\hat{Y}$. Since $\hat{Y} \perp \hat{\varepsilon}$, we also have $\hat{\beta} \perp \hat{\sigma}$.

□

Gauss-Markov Assumptions

We now turn to the case where the errors are iid but not normally distributed. In this case, OLS no longer coincides with MLE so we must consider the use of some central limit theorem to derive asymptotic normality. In particular, we will use the Lindernberg-Feller CLT:

Theorem 3.4 Lindenberg-Feller Central Limit Theorem

Suppose $\{x_{ni}\}$ is a triangular array of $p \times 1$ random vectors such that $z_n = \frac{1}{n}\sum_{i=1}^n x_{ni}$ and

$$V_n = \frac{1}{n}\sum_{i=1}^n \text{Var}(X_{ni}) \to V, \text{ where } V \text{ is psd.}$$

If for every $\varepsilon > 0$ we have

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left\{ ||x_{ni}||^2 \mathbf{1}\left(||x_{ni}|| \geq \varepsilon\sqrt{n} \right) \right\} \to 0,$$

then $\sqrt{n} z_n \xrightarrow{d} \mathcal{N}(0, V)$. Or equivalently, $\sqrt{n} V_n^{-1/2} z_n \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$.

For the asymptotic approximation to hold, we need some weak conditions on the explanatory variables $\mathbf{x}_i$:

- The smallest eigenvalue of $X^T X = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$, namely $\lambda_{\min,n}$, converges to ∞.
- $\max_j P_{jj} = \max_j \mathbf{x}_j^T \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T\right)^{-1} \mathbf{x}_j$ converges to zero.

The first condition states that increasing n always yields more information, while the second condition prohibits any x_j from dominating the others.

Theorem 3.5 Asymptotic normality of coefficients under Gauss-Markov

If the errors ε_i are i.i.d. with mean 0 and variance σ^2, and if $(\mathbf{x}_i)$ satisfies the conditions just given, then the LS estimators $\widehat{\beta}$ are consistent (for β), and the distribution of

$$\left(X^T X\right)^{1/2} (\widehat{\beta} - \beta)$$

converges weakly to $\mathcal{N}_p\left(\mathbf{0}, \sigma^2 I\right)$.

We will proof a slightly easier version of the above statement, by assuming that $x_i^T x_i$ is bonded instead of the second assumption.

Proof. **Consistency.**

The i-th component $\widehat{\beta}_i$ is unbiased and has variance $\sigma^2 \left(\left(X^T X\right)^{-1}\right)_{ii}$, which converges to zero by the first assumption. Then consistency follows from Chebyshev's inequality.

Asymptotic normality.

Suppose $y = X\beta + \omega$ where $\omega_i \sim F$ i.i.d with $\mathbb{E}[\omega_i] = 0, \mathrm{Var}(\omega_i) = \sigma^2$. Suppose further that

$$\frac{1}{n} X^T X \to \Sigma, \text{ where } \Sigma \text{ is psd}$$

Let x_i denote the fixed $p \times 1$ vector of covariates of observation i and assume there exists $M \in \mathbb{R}$ such that $||x_i|| < M < \infty, \ \forall i$.

Step 1: Rearrange OLS estimate

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X\beta + \omega) \\ &= \beta + (X^T X)^{-1} X^T \omega\end{aligned}$$

We can write

$$X^T \omega = \sum_{i=1}^n x_i \omega_i$$

That is, our estimate $\hat{\beta}$ is the true value plus some function of the noise term. While ω_i are iid, premultiplying with x_i yields independently but not identically distributed RVs. We wish to show that this noise term is asymptotically normal.

Step 2: Check Lindenberg-Feller Condition

We now consider the Lindenberg-Feller CLT.

A triangular array is given by random variables in the form

$$\begin{array}{llll} x_{11} \curvearrowleft x_i\omega_i & & & \\ x_{21} & x_{22} & & \\ x_{31} & x_{32} & x_{33} & \\ \vdots & & \dots & \end{array}$$

where the RVs in each row are independent, have expected value zero, and have finite variances.

We consider each element to be an element of $x_i\omega_i$. We can see that the conditions hold as $x_i\omega_i \perp x_j\omega_j$ since ω_i iid. and x_i, x_j are assumed to be fixed, $\mathbb{E}(x_i\omega_i) = x_i\mathbb{E}(\omega_i) = 0$ and $\text{Var}(x_i\omega_i) = \sigma^2 x_i x_i^T < \infty$ by assumption.

Thus, $x_i\omega_i$ is a triangular array with

- row-wise means: $z_n = \frac{1}{n}\sum x_i\omega_i = \frac{1}{n}X^T\omega$.
- row-wise average variance:

$$V_n = \frac{1}{n}\sum_i \text{Var}\left(x_i\omega_i\right) = \frac{1}{n}\sum \sigma^2 x_i x_i^T = \frac{\sigma^2}{n}X^TX \to \sigma^2\Sigma \tag{3.5}$$

  where convergence follows by assumption.

We now turn to verify the LF-condition. Let $\varepsilon > 0$, recall that we assume existence of M such that $||x_i|| < M$, $\forall i$.

$$\begin{aligned} &\frac{1}{n}\sum_i E\left\{\|x_i\omega_i\|^2 \mathbf{1}\left(\|x_i\omega_i\| \geqslant \varepsilon\sqrt{n}\right)\right\} \\ &= \frac{1}{n}\sum_i E\left\{\omega_i^2 \|x_i\|^2 \mathbf{1}\left(|\omega_i|\,\|x_i\| \geqslant \varepsilon\sqrt{n}\right)\right\} \\ &< \frac{1}{n}\sum_i E\left\{\omega_i^2 M^2 \mathbf{1}\left(|\omega_i|\, M \geqslant \varepsilon\sqrt{n}\right\}\right. \\ &= E\left\{\omega^2 M^2 \mathbf{1}(|\omega| M \geqslant \varepsilon\sqrt{n})\right\} \\ &=: E\left\{T_n\right\} \end{aligned}$$

Note that $T_n \xrightarrow{p} 0$ as $\varepsilon\sqrt{n}M^{-1} \to \infty$ as ω are iid with finite variance.[4]

Define $Z = M^2\omega^2, \mathbb{E}\left[Z\right] = M^2\sigma^2 < \infty$. Note that $\forall n \in \mathbb{N} : T_n \leq Z$. Since T_n converges pointwise to the zero function and we have an integrable function which dominates T_n we can apply the *Dominated convergence theorem* to deduce convergence of $\mathbb{E}(T_n)$:

$$T_n \xrightarrow{p} 0 \implies T_n \xrightarrow{d} 0 \implies \mathbb{E}\left[T_n\right] \to 0$$

[4] If ω_i are iid with finite second moment, we have from Chebyshev's inequlity that

$$\mathbb{E}\left[\omega_i^2 1_{[|\omega_i|>d]}\right] = \mathbb{E}\left[\omega_1^2 1_{[|\omega_1|>d]}\right] \xrightarrow{d\to\infty} 0.$$

3: Putting it all together

Recall:

$$\hat{\beta} = \beta + (X^T X)^{-1} X^T \omega$$
$$z_n = \frac{1}{n} \sum x_i \omega_i = \frac{1}{n} X^T \omega$$
$$V_n = \frac{1}{n} \sum \mathrm{Var}\left(x_i \omega_i\right) \to \sigma^2 \Sigma$$

Then the LF-CLT says

$$\sqrt{n} z_n \xrightarrow{d} \mathcal{N}(0, V)$$

Hence,

$$\frac{1}{\sqrt{n}} X^T \omega \xrightarrow{d} \mathcal{N}(0, \sigma^2 \Sigma)$$

We know that $(X^T X)(\hat{\beta} - \beta) = X^T \omega$, thus (dividing by the square root of the variance from (3.5)), we normalise and obtain

$$\frac{1}{\sigma} (X^T X)^{-1/2} (X^T X)(\hat{\beta} - \beta) = \frac{1}{\sigma} (X^T X)^{1/2} (\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbb{I})$$

□

3.3 Testing

We next prove the distribution of various potential test statistics. Note that we always assume $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_{n \times n})$. In particular, p-values in summary outputs in R rely on this normality assumption. If instead only Gauss-Markov assumptions hold, p values and confidence intervals are only valid asymptotically.

Confidence interval for β

Proposition 3.2

Assume $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_{n \times n})$ then $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$. Further, let $\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n-p}$. We have $\frac{\sum \hat{\varepsilon}_i^2}{\sigma^2} \sim \chi^2_{n-p}$ and $\hat{\sigma} \perp \hat{\beta}$. Then:

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma} \sqrt{\left((X^T X)^{-1}\right)_{ii}}} \sim t_{n-p}$$

Proof. Recall, by definition of the t-distribution that if $Z \sim \mathcal{N}(0,1), V \sim \chi^2_n, Z \perp V$, then

$$T = \frac{Z}{\sqrt{V/n}} \sim t_n$$

Since $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2 \left((X^T X)^{-1}\right)_{ii})$, we have

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\left((X^T X)^{-1}\right)_{ii}}} \sim \mathcal{N}(0,1)$$
$$V = \frac{\sum \hat{\varepsilon}_i}{\sigma^2} \sim \chi^2_{n-p} \text{ and } V \perp Z$$

Thus

$$T = \frac{\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{\left((X^T X)^{-1}\right)_{ii}}}}{\sqrt{\frac{\sum \hat{\varepsilon}_i^2}{\sigma^2(n-p)}}} = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{\left((X^T X)^{-1}\right)_{ii}}} \sim t_{n-p}.$$

Here we used $\sum \hat{\varepsilon}_i^2/(n-p) = \hat{\sigma}^2$. □

Using our typical 'Wald-type-CI', we can achieve a α level CI by rewriting

$$\hat{\beta}_i - q_{t_{\alpha/2,n-p}} \hat{\sigma}_{\hat{\beta}_i} \leq \beta_i \leq \hat{\beta}_i + q_{t_{\alpha/2,n-p}} \hat{\sigma}_{\hat{\beta}_i}$$

where $\hat{\sigma}_{\hat{\beta}_i} = \hat{\sigma}\sqrt{\left((X^T X)^{-1}\right)_{ii}}$.

Confidence Interval for $\mathbb{E}(Y)$

The confidence interval for the expected value of Y (i.e., confidence interval for the level of the regression line at a point), is entirely similar to above. Noting that

$$\hat{Y}_i \sim \mathcal{N}(\mu_i, \sigma^2 H_{ii})$$

where H is the matrix used in $\hat{y} = Hy$, we obtain the test statistic

$$T = \frac{\hat{y}_i - \mu_i}{\hat{\sigma}\sqrt{H_{ii}}} \sim t_{n-p}.$$

The proof of the distribution follows our approach above.

Prediction interval

If we want to have a confidence interval for the predicted interval of accuracy single observation, i.e., a **prediction interval**, we need to consider not only the source of uncertainty from a wrong estimate of β but also the underlying noise in the data. Thus, prediction intervals are larger than the corresponding CI for $\mathbb{E}(\hat{Y})$.

$y_0 \perp \hat{y}_0$, since we consider a new observation $y_0 = x_0^T \beta + \varepsilon_0$ which was not used in the estimation of $\hat{\beta}$.

Consider $D = y_0 - \hat{y}_0 \sim \mathcal{N}(\mu_D, \sigma_D^2)$ where normality follows as this is a difference between two independent normal variables. By unbiasedness and $\mathbb{E}(\varepsilon) = 0$, $\mu_D = 0$. By independence[5]

$$\begin{aligned}\sigma_D^2 &= \text{Var}(y_0) + \text{Var}(\hat{y}_0) \\ &= \sigma^2(1 + x_0^T (X^T X)^{-1} x_0))\end{aligned}$$

[5] As $n \to \infty$, the second term goes to zero while the first term will always remain non-zero.

Continuing as in our derivation above, we arrive at the test statistic

$$T = \frac{y_0 - \hat{y}_0}{\hat{\sigma}\sqrt{1 + x_0^T (X^T X)^{-1} x_0}} \sim t_{n-p}$$

Note that a prediction interval is interested in *one new observation.* If we want to arrive at an interval that contains $P\%$ of *all future observations* (with a certain level of confidence), we must instead consider a **tolerance interval.**

3.4 F-Test

Colinearity

Highly colinear variables lead to two problems:

- Interpretation is difficult since these variables tend to move together. Holding one fixed is not a natural phenomenon.
- Reduced accuracy in paramter estimates

We may measure colinearity using **Variance Inflation Factor (VIF)**:

$$\text{VIF} = \frac{1}{1-R_j^2} \text{ where } R_j^2 \text{ is the } R^2 \text{ value of the regression } x_j \sim x_{-j}$$

I.e., how well can x_j be explained by the other covariates? In practice, t-tests tend not to be significant but become significant when other variables are dropped. This will motivate the use of tests for whether at least one of multiple coefficients is non-zero.

Recap: Different likelihood-based tests

Recall that we can perform inference test based on the likelihood function in three different ways:

- *Wald type:* differences between $\hat{\alpha}$ and α_0 (closely related to t-test)[6]
- *LR:* Difference in likelihood $\log L(\hat{\alpha})$ and $\log L(\alpha_0)$ (closely related to F-test)[7]
- *Score*: Slope of derivative of likelihood w.r.t. α_0

[6] Tests for individual coefficients. t-Test uses a t-distribution, square root of Wald test uses asymptotically normal distribution. Equivalent as $n \to \infty$.

[7] LR uses ratio of likelihood functions, F-test uses ratio of RSS. It turns out that for linear regression, the (partial) F-test is equivalent to the likelihood ratio test as RSS $\propto$ likelihood.

General form of F-tests

We focus on F-Tests. We are interest in asking questions such as are two coefficients identical, or are multiple coefficients equal to 0? These questions take the general form:[8]

$$H_0 : B\beta = \text{b for some } ((p-q)\times p \text{ })\text{-matrix } B \text{ and some } (p-q)\text{-vector } b \text{ (often } b=0)$$

We can perform one **partial F-Test** or multiple t-tests (one per row, with multiple testing adjustments). Partial F-tests compare nested models whereby the large model consists of p parameters, thh small of q parameters and therea are $p-q$ side constraints.

Example 3.1

Suppose we want to test the null hypothesis $\beta_3, \beta_4, \beta_5 = 0$. Then we may write this in matrix form as follows:

$$\begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Example 3.2 (Global F-Test)

We test the null hypothesis that all coefficients other than the intercept are 0. We have

$$B = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}; \quad \mathbf{b} = \mathbf{0}$$

The test statistic is then $\sim F_{p-1,n-p}$ since we have $q = 1$ (intercept in restricted models).
As under the null, the simple model always predicts $\bar{Y}$, we can rewrite this test statistic using the relation between the F-test and SSE as follows:

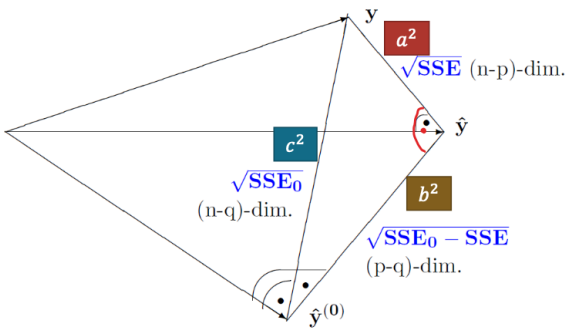
$$\begin{aligned} F &= \frac{n-p}{k} \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum \hat{\varepsilon}_i^2} \\ &= \frac{n-p}{k} \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2 - \sum (\hat{y}_i - \bar{y})^2} \\ &= \frac{n-p}{k} \frac{R^2}{1-R^2}. \end{aligned}$$

Example 3.3 (Exact F-Test for specific value of β)

If we specify a complete set of values for the vector β under the null hypothesis, our

[8] A particularly common F-test is the global F-test which tests whether at least one coefficient other than the intercept is different from 0 (vs all coefficients being zero).

Figure 4: Graphical Intuition for F-Test



test statistic reduces to

$$\frac{(\widehat{\beta}-\beta)^T X^T X(\widehat{\beta}-\beta)}{p\widehat{\sigma}^2} \sim F_{p,n-p}$$

Intuition Partial F-Test

Suppose we have a point y and we first project it onto a p-dimensional subspace to obtain $\hat{y}$. Next consider adding additional side-constraints ($p-q$ such constraints), and projecting onto that subspace to obtain $\hat{y}^0$. Note that $\hat{y}^0$ is also in the column space of X. The line connecting $\hat{y}^0$ and $\hat{y}$ must thus be orthogonal to the line from y to $\hat{y}$. We therefore have a right-angled triangle in Figure 4 and can use Pythagoras to obtain a relation between the different line lengths. An intuitive test statistic would be to say that the simple model is true if b^2 is small relative to a^2.

Under the assumption that the simple model is true, we have

$$\frac{(SSE_0 - SSE)/(p-q)}{SSE/(n-p)} = \frac{\left\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{(0)}\right\|^2/(p-q)}{\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2/(n-p)} \sim F_{p-q,n-p}$$

To show this, we use the following lemma which can be obtained using Lagrange multipliers:

Lemma 3.1

The least squares estimator $\widehat{\beta}_{(0)}$ under the supplementary condition $B\beta = \mathbf{b}$ is

$$\widehat{\beta}_{(0)} = \widehat{\beta} - \left(X^T X\right)^{-1} B^T \left(B\left(X^T X\right)^{-1} B^T\right)^{-1} (B\widehat{\beta} - \mathbf{b})$$

Furthermore,[a]

$$\underbrace{SSE_0}_{c^2} = \underbrace{SSE}_{a^2} + \underbrace{(B\widehat{\beta}-\mathbf{b})^T \left(B\left(X^T X\right)^{-1} B^T\right)^{-1} (B\widehat{\beta}-\mathbf{b})}_{b^2}.$$

[a] To obtain this, one may use

$$b^2 = \left\|\hat{Y} - \hat{Y}_0\right\|^2 = \left(X\left(\hat{\beta} - \hat{\beta}_0\right)\right)^T X\left(\hat{\beta} - \hat{\beta}_0\right) = \cdots$$

We use this, and the fact $\hat{\sigma}^2 = \varepsilon^T\varepsilon/(n-p)$, to rewrite:

$$\frac{(SSE_0 - SSE)/(p-q)}{SSE/(n-p)} = \frac{(B\widehat{\beta} - \mathbf{b})^T \left(B\left(X^TX\right)^{-1}B^T\right)^{-1}(B\widehat{\beta} - \mathbf{b})}{(p-q)\widehat{\sigma}^2} \sim F_{p-q,n-p}$$

One can show that this is a likelihood ratio test. We now show that this follows an F-distribution if $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{1})$.

Proof. Recall that if $Y \sim \mathcal{N}(\mu, \Sigma)$, then $(y-\mu)^T\Sigma^{-1}(y-\mu) \sim \chi^2_n$. Also, $X = \frac{S_1/d_1}{S_2/d_2} \sim F_{d_1,d_2}$ if $S_i \sim \chi^2_{d_i}$ for $i = 1, 2$ and $S_1 \perp S_2$.

Since $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$,

$$B\hat{\beta} \sim N(\underbrace{B\beta}_{=b}, \sigma^2 B(X^TX)^{-1}B^T)$$

Thus

$$S_1 := (B\hat{\beta} - b)^T\left(\sigma^2 B(X^TX)^{-1}B^T\right)^{-1}(B\hat{\beta} - b) \sim \chi^2_{p-q}.$$

We obtain the degrees of freedom by noting that b is a $(p-q)$-dimensional vector.

We have previously shown in Proposition 3.1 (which relies on $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{1})$) that

$$\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{\sigma^2} \sim \chi^2_{n-p}$$

A simple rearrangement, using $\hat{\sigma}^2 = \sum \hat{\varepsilon}_i^2/(n-p)$ yields

$$S_2 := \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Since $\hat{\varepsilon} \perp \hat{\beta}$, we have $S_1 \perp S_2$.

Finally, after cancelling out some terms,

$$\frac{S_1/(p-q)}{S_2/(n-p)} = \frac{(B\hat{\beta} - b)^T\left(B(X^TX)^{-1}B^T\right)^{-1}(B\hat{\beta} - b)}{(p-q)\hat{\sigma}^2} \sim F_{p-q,n-p}.$$

□

In R: Fit two nested models (e.g. fm and fm2) and compare with anova(fm, fm2).

3.5 Residual Analysis

Recall that errors are not identical to residuals. In particular,

- **Errors**: $\varepsilon \sim N\left(0, \sigma^2 1\right) \rightarrow$ Errors are uncorrelated and have constant variance
- **Residuals**: $\hat{\varepsilon} \sim N\left(0, \sigma^2 M\right) \rightarrow$ Residuals are correlated and have different variance $\operatorname{Var}\left(\hat{\varepsilon}_i\right) = \sigma^2(1- H_{ii})$
- By design: $\hat{y}$ and $\hat{\varepsilon}$ are uncorrelated
- We sometimes instead consider standardized residuals: $\hat{\varepsilon}_i^S = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-H_{ii}}} \rightarrow$ unit variance if the true model is correct.[9]

There are four main assumptions to check:

1. Independent samples:
   - Plot residuals vs recording time or order of observation: is there serial correlation or clustering?
2. Functional relationship and constant variance
   - Scatter plot $x - y$ in simple regression
   - Tukey-Anscombe plot (residual vs fitted value $\hat{\varepsilon} - \hat{y}$)[10]
   - Scale Location plot (standardised residual vs fitted value with horizontal smoother), see Figure 5
3. Normal distribution of errors
   - QQ-Plot of residuals

Outliers

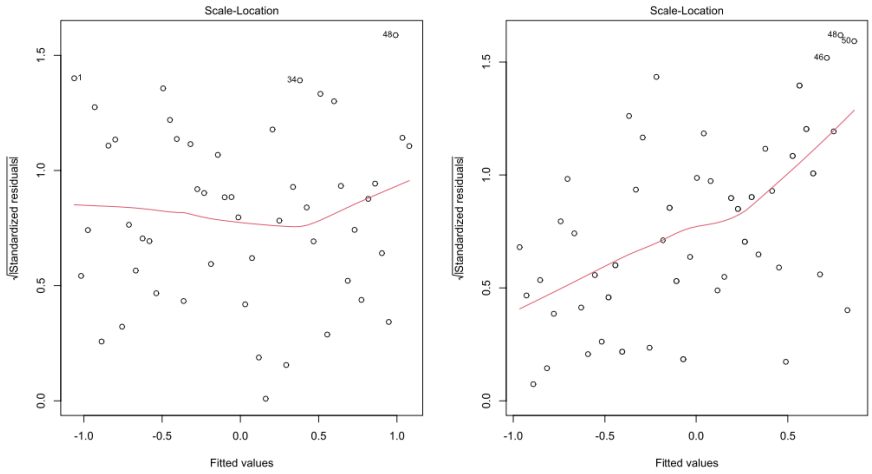
Outliers can occur in the y or x direction. Some ways of measuring outliers are:

- **Leverage** H_{ii}: How influential is y_i on prediction $\widehat{y_i}$?
  - $\widehat{Y} = HY \rightarrow \frac{d\widehat{y_i}}{dy_i} = H_{ii}$ depends on X but not on y
- **Cook's Distance** $\mathbf{D}_i$: How influential is y_i on whole fit?
  - Note: $\hat{y}_{j(i)}$ is the fitted response value for observation j obtained when excluding observation i

[9] If for instance $\operatorname{Var}(\epsilon) = \Sigma \neq \sigma^2 \mathbb{I}$, then the standardization will not yield a unit variance since we will not have $\hat{\varepsilon} \sim N\left(0, \sigma^2 M\right)$. We can spot this in adequate plots.

[10] Note: cannot be too picky with varying variances since residuals do not have constant variances.

Figure 5: Examples of scale-location plots. The left plot is okay, the right hand plot indicates problematic violations of assumptions as the variance seems to increase with the value of $\hat{y}$



- $D_i = \sum_{j=1}^{n} \frac{\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{p\hat{\sigma}^2}$ can be transformed to $D_i = \frac{\hat{\varepsilon}_i^{\,2}}{p\widehat{\sigma}^2} \cdot \left(\frac{H_{ii}}{(1-H_{ii})^2}\right)$[11]
- Rule of thumb: $D_i > 1$ is problematic

We can investigate outliers in standard residual plots (e.g.„ residual vs leverage).

4 Model Selection

If we aim at **inference**, we should postulate one model as this allows for valid p-values.

If we instead explore multiple models, p-values etc. are no longer valid, we have a multiple testing issue. This is a research topic, post-selection inference.

Consequences of too many or too few variables in the model may be:

1. *Too few* (sparse model): Bias, but reduced variance[12]
2. *Too many*: Unbiased, large variance

In particular, overfitting is an issue which can reduce prediction accuracy when considering bigger models. This is the fundamental issue behind the bias-variance trade-off. We have

$$\text{MSE} = \text{Var} + \text{Bias}^2 .$$

[11] Similarly, there is a transformation which can be applied to calculate LOOCV errors: $CV = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_{i(i)}^2$ can be transformed to $CV = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\hat{\varepsilon}_i}{1-H_{ii}}\right)^2$

[12] The model is unbiased if the missing variables are not correlated with the included Xs. E.g., if we have a completely orthogonal design matrix.

For this, note that when z is an estimator of the unknown constant c,

$$E\left[(z-c)^2\right] = E\left[((z-E[z])+(E[z]-c))^2\right]$$
$$= \underbrace{E\left[(z-E(z))^2\right]}_{\text{Var}(z)} + 2E\Big[\underbrace{(z-E[z])}_{\to 0} \cdot \underbrace{(E[z]-c)}_{\text{constant}}\Big] + \underbrace{E\left[(E[z]-c)^2\right]}_{\text{bias}^2}$$

Note that $E[z]-c$ is constant since we take expectation with respect to z, so c is a constant and $E(E(z))$ is also just the expectation of a constant real number.

4.1 Approaches for model selection

Local approaches using F and t-tests

Search strategies based on partial F-Test:

- Stepwise forward
- Stepwise backward (only possible if $n > p$)
- Add "most significant" / remove "least significant" variable according to partial F-test

Drawbacks:

- p-values are used but not strictly valid anymore
- Only nested models can be compared

Global Approaches using Cp and AIC

Often we do not want to use data splitting or cross-validation due to computational concerns. In this case, we can **estimate Test MSE** by using a formula that approximates the additional error in test MSE out-of-sample. That is, we want to minimize a quantitiy like

$$\min RSS + c \cdot p$$

where we have a penalty for complex models.

- Mallow's C_p

$$C_p = \frac{RSS}{\hat{\sigma}^2} + 2|M| - n$$

- Akaike's Information Criterion:

$$AIC = -2 \cdot l\left(\hat{\theta}_M\right) + 2 \cdot p$$

- Bayesian Information Criterion:

$$BIC = -2 \cdot l\left(\hat{\theta}_M\right) + \log(n) \cdot p$$

The last two criteria are more general than Mallow's C_p as they are valid for all models which admit a likelihood.

4.2 Derivation of Mallow's C_p

We want to derive Mallow's C_p.

Setup:

- Assume $y_i, i = 1, \ldots, n$ independent.
- $\mathbb{E}[y_i] = \mu_i, \text{Var}(y_i) = \sigma^2$
- Regressors $1, x_1, \ldots, x_k \implies X$ is a $n \times (k+1)$ design matrix
- Consider a subset $M \subseteq \{0, 1, \ldots, k\}$, denote the design matrix of this reduced set of regressors by X_M which is $n \times |M|$

Then

$$\hat{\beta}_M = \left(x_M^T X_M\right)^{-1} X_M^T y, \quad \hat{y}_M = X_M \hat{\beta}_M = H_M y$$

Facts:

1. $E\left(\hat{Y}_M\right) = H_M E(y) = X_M \left(X_M^\top X_M\right)^{-1} X_M^\top X\beta.$
   - Note that we may have bias as the X_M and X term do not cancel each other out
2. $\text{Cov}(\hat{y}_M) = \sigma^2 H_M$
3. $\sum_{i=1}^n \text{Var}\left(\hat{Y}_{iM}\right) = \text{tr}\left(\sigma^2 H_M\right) = \sigma^2 \,\text{tr}\left(X_M \left(X^\top X_M\right)^{-1} X_M^\top\right) = \sigma^2 |M|$
4. Sum of mean squared errors:[13]
   Use $\mu_{iM} = \mathbb{E}[\hat{y}_{iM}]$. Then

$$\begin{aligned}
\text{SMSE} =& \sum_{i=1}^n \mathbb{E}\left[(\hat{y}_{iM} - \mu_i)^2\right] \\
=& \sum_{i=1}^n E\left((\hat{y}_{iM} - \mu_{im}) + (\mu_{im} - \mu_i)\right)^2 \\
=& \sum \text{Var}(\hat{y}_{iM}) + 2\sum E(\underbrace{(\hat{y}_{iM} - \mu_{im})}_{\to 0}(\mu_{iM} - \mu_i)) \\
&+ \sum \underbrace{E(\mu_{iM} - \mu_i)^2}_{\text{constant, can remove } E} = \\
=& \sigma^2 |M| + \sum (\mu_{iM} - \mu_i)^2
\end{aligned} \tag{4.1}$$

   We can then define

$$\gamma(M) = \frac{\text{SMSE}}{\sigma^2} = |M| + \frac{\text{bias}^2}{\sigma^2} \tag{4.2}$$

   We note that $\gamma(M) \geq M$ with equality if the model is unbiased.

[13] Note, this is a bit of a misnomer. Instead it is the sum of expected squared deviations between predictions and expected value for y_i.

5. Future observations:

   Assume that we record again observations y for all $i = 1, \dots, n$ with the same regressors $x_{i1}, \dots, x_{ik}$. Then

$$Y_{n+i} = \mu_i + \underbrace{\varepsilon_{n+i}}_{\text{independent of} \varepsilon_1, \dots, \varepsilon_n}$$

   Since regressors are identical, our prediction $\hat{Y}_{n+i} = \hat{y}_n$.

   We define the *expected squared prediction error (SPSE:)*[14]

$$\begin{aligned}
\text{SPSE} &= \sum_{i=1}^{n} \mathbb{E}\left[\left(Y_{n+i} - \hat{Y}_{iM}\right)^2\right] \\
&= \sum E\left((Y_{n+i} - \mu_i) - \left(\hat{Y}_{im} - \mu_i\right)\right)^2 \\
&= \sum (E\left(Y_{n+1} - \mu_i\right)^2 - 2E\left(\underbrace{\left(Y_{n+i} - \mu_i\right)\left(\hat{Y}_{iM} - \mu_i\right)}_{\to 0}\right) + E\left(\hat{Y}_{im} - \mu_i\right)^2) \quad (4.3) \\
&= n\sigma^2 + SMSE \\
&= \underbrace{n\sigma^2}_{(1)} + \underbrace{|M|\sigma^2}_{(2)} + \underbrace{\sum \left(\mu_{iM} - \mu_i\right)^2}_{(3)}
\end{aligned}$$

   Note in the third line that y_{n+i} and $\hat{y}_{iM}$ are independent. Therefore the expectation of the product is the product of expectations and we know that $E(Y_{n+i} - \mu_i) = 0$.

   We have the following three elements: (1) Irreducible error, (2) Variance, (3) Bias.

Before deriving Mallow's C_p, we show the following relation:

Lemma 4.1

$$\mathbb{E}(\text{RSS}) = \text{ SPSE } - 2|M|\sigma^2$$

For this, note that $\mathbb{E}(RSS)$ is the sum of expected squared residuals in sample (as opposed to SPSE which considers out of sample residuals).

Proof.

$$\begin{aligned}
E(RSS) &= E\left(\sum \left(y_i - \hat{y}_i^M\right)^2\right) = \sum E\left(\left(y_i - \hat{y}_i^M\right)^2\right) \\
&= \sum \underbrace{\text{Var}\left(y_i - \hat{y}_i^M\right)}_{(1)} + \underbrace{\sum \left(E\left(y_i - \hat{y}_i^M\right)\right)^2}_{(2)}
\end{aligned}$$

which follows from the usual formula for the variance, $\text{Var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

ad (1):

[14] Note that we only consider squared errors on new observations.

$$y - \hat{y}^M = y - H^M y = \left(\mathbf{1} - H^M\right) y$$
$$\operatorname{Cov}\left(\hat{y} - \hat{y}^M\right) = \left(\mathbf{1} - H^M\right) \sigma^2$$

Therefore,

$$\sum \operatorname{Var}\left(y_i - \hat{y}_i^M\right) = \sigma^2 \operatorname{tr}\left(\mathbf{1} - H^M\right)$$
$$= \sigma^2 \left(n - \operatorname{tr}\left(H^M\right)\right) = \sigma^2 (n - |M|)$$

ad (2):

$$\sum \left(E\left(y_i - \hat{y}_i^M\right)\right)^2 = \sum \left(E\left(\hat{y}_i^M\right) - \mu_i\right)^2 =$$
$$= \text{SMSE} - |M|\sigma^2$$
$$= \text{SPSE} - n\sigma^2 - |M|\sigma^2$$

In the first step we use $\mathbb{E}(y_i) = \mu_i$. Then we used equations (4.1). In particular, we use $E(\hat{y}_{iM}) = \mu_{iM}$ and the last line in that block of equations. Finally, we use (4.3) to get the result.

Continuing with what we want to show, we have

$$\mathbb{E}\left[RSS\right] = (n - |M|)\sigma^2 + SPSE - n\sigma^2 - |M|\sigma^2 =$$
$$= SPSE - 2\left|M\right|\sigma^2$$

□

Mallow's C_p now wants to estimate $\gamma(M) = \frac{SMSE}{\sigma^2}$ as defined in (4.2). From the above,

$$\begin{aligned} \text{SPSE} &= n\sigma^2 + SMSE = E(RSS) + 2|M|\sigma^2 \\ \Rightarrow \text{SMSE} &= E(RSS) + 2|M|\sigma^2 - n \cdot \sigma^2 \end{aligned}$$

We now replace unobserved aspects with estimates quantities,

$$C_p = \frac{RSS}{\hat{\sigma}^2} + 2|M| - n$$

4.3 Relation between C_p and AIC for Linear Regression

In linear regression,[15]

$$\text{AIC} = -2 \cdot l\left(\hat{\beta}_M, \hat{\sigma}^2\right) + 2(|M| + 1)$$

The term of the log-likelihood is given by

$$-2l\left(\hat{\beta}_M, \hat{\sigma}^2\right) = n \log\left(\hat{\sigma}^2\right) + \frac{1}{\hat{\sigma}^2}\left(\mathbf{y} - \mathbf{X}_M \hat{\beta}_M\right)'\left(\mathbf{y} - \mathbf{X}_M \hat{\beta}_M\right)$$
$$= n \log\left(\hat{\sigma}^2\right) + \frac{n\hat{\sigma}^2}{\hat{\sigma}^2}$$
$$= n \log\left(\hat{\sigma}^2\right) + n.$$

[15] Note that we have $|M|+1$ parameters since we estimate $|M|$ coefficients and the residual standard error.

We get

$$\text{AIC} = n \cdot \log\left(\hat{\sigma}^2\right) + 2(|M| + 1) + n \tag{4.4}$$

This looks quite different to C_p, but we can show similarity by applying a Taylor expansion of $\log \hat{\sigma}^2$ around σ^2 to obtain

$$\log \hat{\sigma}^2 \approx \log \sigma^2 + \frac{1}{\sigma^2}\left(\hat{\sigma}^2 - \sigma^2\right) = \log \sigma^2 + \frac{SSE}{n\sigma^2} - 1$$

Thus, ignoring the constant $2 + n$ in (4.4),

$$AIC \approx n \log\left(\sigma^2\right) + \underbrace{\frac{SSE}{\sigma^2} - n + 2\,|M|}_{\approx C_p}$$

4.4 General Search Strategies

- **Exhaustive Search**
  - For given $p = 1, \ldots, p_{\max}$: Find best model according to RSS[16]
  - We obtain $p_{\max}$ different models
  - We then choose the best one with C_p or AIC,
    - Pros: Finds global optimum
    - Cons: Computational cost p variables $\rightarrow 2^p$ subsets
- **Forward / Backward Selection**
  - Forward: Start with empty model, add one variable at a time
  - Backward: Start with full model, remove one variable at a time
  - E.g., in forward selection we calculate C_p for each potential $p' + 1$ variable model with first p variables fixed after having found best p' model. If there is no model with a better C_p value, we stop.[17]
    - Pros: Fast
    - Cons: Only local optimum $\rightarrow$ Forward and Backward often get to different final models

In R:

```
regsubsets(y ~ ., data, method = "exhaustive", nvmax = 15)
```

[16] We do not require an information criteria for a fixed number of parameters p since in this case the model with lowest RSS automatically has the lowest value for the information criteria.

[17] The R implementation in regsubsets finds the best model according to RSS for any potential number of variables and only in the end chooses according to information criteria between models with different number of parameters.

5 Non i.i.d. errors

Assume $\text{Cov}(\varepsilon) = \sigma^2 \mathbb{I}$. Then we can show that

$$\text{Cov}\left(\hat{\beta}\right) = \sigma^2 (X^T X)^{-1}$$

Now, instead assume $\text{Cov}(\varepsilon) = \sigma^2 W^{-1}$. Then

$$\text{Cov}\left(\hat{\beta}\right) = \sigma^2 (X^T X)^{-1} X^T W^{-1} X (X^T X)^{-1} \neq \sigma^2 (X^T X)^{-1}$$

$\implies$ Using 'normal' OLS leads to wrong p-values and CIs.

Often times we can detect deviations from $\text{Cov}(\varepsilon) = \sigma^2 \mathbb{I}$ by investigating TA-plots or from context knowledge (e.g., grouped measurements). In particular if we have additional information about the form of the covariance matrix, we can often use such information to estimate the covariance matrix and obtain correct CIs.

5.1 Known covariance matrix – Generalised Least Squares

Suppose we know the covariance matrix of our errors up to a factor. E.g.,

$$\mathbf{Y} = X\beta + \varepsilon \quad \varepsilon \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \Sigma\right)$$

where Σ is known and PD. By Proposition 1.11 (and due to PD, thus invertibility), there exists an invertible matrix A such that

$$\Sigma^{-1} = (AA^T)^{-1}$$

We can thus formulate our OLS problem in a 'tilde' dimension where we use

$$\tilde{\mathbf{Y}} := A^{-1}\mathbf{Y} = A^{-1}(X\beta + \varepsilon) = \underbrace{A^{-1}X}_{\tilde{X}}\beta + \underbrace{A^{-1}\varepsilon}_{\tilde{\varepsilon}} = \tilde{X}\beta + \tilde{\varepsilon}$$

We have

$$\begin{aligned}
\mathbb{E}\tilde{\varepsilon} &= \mathbb{E}A^{-1}\varepsilon = A^{-1}\mathbb{E}\varepsilon = \mathbf{0} \\
\text{Cov}[\tilde{\varepsilon}] &= \text{Cov}\left[A^{-1}\varepsilon\right] = A^{-1}\,\text{Cov}[\varepsilon]\left(A^{-1}\right)^T \\
&= A^{-1}\sigma^2\left(AA^T\right)\left(A^{-1}\right)^T = \sigma^2 I.
\end{aligned}$$

We can thus apply our usual OLS results to estimate $\hat{\beta}$ from the 'tilde model'

$$\|\tilde{\mathbf{Y}} - \tilde{X}\beta\|^2 = (\mathbf{Y} - X\beta)^T A^{-T} A^{-1} (\mathbf{Y} - X\beta) = (\mathbf{Y} - X\beta)^T \Sigma^{-1} (\mathbf{Y} - X\beta).$$

We can think of this as rescaling our residuals such that observations with larger variance (more uncertainty) receive a lower weight. This will be more obvious in the special case where Σ is diagonal which we treat below.

Solving the above for $\hat{\beta}$ gives us our usual result in the tilde dimension:

$$\widehat{\beta} = \left(\tilde{X}^T\tilde{X}\right)^{-1}\tilde{X}^T\tilde{\mathbf{Y}} = \left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\mathbf{Y}.$$

In the last step we plug back in our definitions of $\tilde{X}, \tilde{Y}$. It is then easy to see that

$$\widehat{\beta} \sim \mathcal{N}_p\left(\beta, \sigma^2\left(X^T\Sigma^{-1}X\right)^{-1}\right), \quad \hat{\sigma}^2 = \frac{1}{n-p}\varepsilon^T\Sigma^{-1}\varepsilon$$

If we know that each y_i is an average of w_i-many i.i.d. measurements, we obtain the special case of **weighted least squares.**

$$\mathrm{Var}\left(\varepsilon_i\right) = \frac{\sigma^2}{w_i}; \quad \mathrm{Cov}(\varepsilon) = \sigma^2\,\mathrm{diag}\left(\frac{1}{w_1}, \ldots, \frac{1}{w_n}\right) = \sigma^2\Sigma$$

Thus

$$\begin{aligned}|\tilde{Y} - \tilde{X}\beta|^2 &= \left[A^{-1}(Y - X\beta)\right]^T A^{-1}\left[Y - X\beta\right] \\ &= (Y - X\beta)^T\Sigma^{-1}(Y - X\beta) = \sum_{i=1}^{n} w_i\left(y_i - x_i^T\beta\right)^2\end{aligned}$$

Each observation in our objective function is weighted proportionally to the **number of samples** or **inverse variance of error**. Note that WLS in general encapsulates cases where we have heteroskedastic but uncorrelated errors and know the heteroskedasticity strucutre. In this case, we have $\Sigma^{-1} = W^{1/2}W^{1/2}$ where $W = \mathrm{diag}(w_1, \ldots, w_n)$, so $A^{-1} = A^{-\top} = W^{1/2}$.

In R: fm <− lm(y ∼ x, data = df, weights = nreps), weights arugment

From the fact that we can rescale our model to a tilde dimension with diagonal covariance matrix, we can apply many of our previous results, in particular:

Proposition 5.1 Gauss-Markov for WLS

Among all linear and unbiased estimators $\hat{\beta}^L = \mathbf{Ay}$, the WLS estimator has minimal variance, i.e.,

$$\mathrm{Var}\left(\hat{\beta}_j\right) \leq \mathrm{Var}\left(\hat{\beta}_j^L\right), \quad j = 0, \ldots, k$$

A similar extension holds with normally distributed errors and WLS being the minimal variance unbiased estimator.

Note that we could use our normal OLS estimate and the fact that

$$\mathrm{Var}(\hat{\beta}_{\mathrm{OLS}}) = \sigma^2(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$$

while plugging in Σ to obtain correct coverage probabilities. However, the above application of the Gauss-Markov theorem shows that instead using Generalized Least Squares leads to more efficient estimates.

5.2 Unknown covariance matrix, known structure

MLE and two-stage OLS

If our covariance matrix is unknown, but we can specify the structure, we can usually use this information to model the covariance matrix. There are two approaches:

1. Two-stage procedure
   (a) Fit OLS
   (b) Use OLS residual estimates and information about Covariance structure to model coefficients of covariance
   (c) Use above to derive weights and fit GLS
   (d) Iterate
   (e) Example: Serial correlation of AR(1) form

$$\text{Cov}(\varepsilon) = \sigma^2 W^{-1} = \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{pmatrix}$$

   We estimate residuals using OLS, then estimate $\rho_1, ..., \rho_k$ using residuals and use ρ to define weights to be used in GLS.

2. Maximum-Likelihood
   (a) Make assumption about distribution of ε and write out likelihood using known information about covariance matrix
   (b) Estimate covariance matrix using MLE

Huber-White HC (Sandwich) Estimator

Consider the model

$$\begin{aligned} &Y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}_n(0, D) \\ &D = \text{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right), \sigma_i^2 = \text{Var}\left(\varepsilon_i\right). \end{aligned}$$

We know that

$$\hat{\beta} \sim \mathcal{N}_p\left(\beta, \left(X^T X\right)^{-1} X^T D X \left(X^T X\right)^{-1}\right).$$

For statistical testing or confidence intervals, we need to estimate the covariance matrix. An easy but powerful approach is to use

$$\hat{D} = \text{diag}\left(r_1^2, \ldots, r_n^2\right)$$

where $r = Y - X\hat{\beta}$ are the residuals (from the least squares estimator). The estimated covariance matrix is then

$$\hat{V} = \widehat{\text{Cov}}(\hat{\beta}) = \left(X^T X\right)^{-1} X^T \hat{D} X \left(X^T X\right)^{-1}\Big).$$

Note that it may seem implausible to estimate this, as we have n observation and the matrix $\hat{D}$ consists of n parameters. However, we can show that we can estimate the term $X^T\hat{D}X$ consistently. In particular, since we know $V^{-1/2}(\hat{\beta}-\beta) \sim \mathcal{N}_p(0,I)$, we now want to show:

$$\hat{V}^{-1/2}(\hat{\beta}-\beta) \Longrightarrow \mathcal{N}_p(0,I) \text{ as } n \to \infty$$

It follows from Slutsky's theorem (Theorem A.1) that this is the case if we can consistently estimate V.[18] Thus, we want to show that $\hat{\beta} \xrightarrow{p} \beta$ as $n \to \infty$ and

$$n(\widehat{\text{Cov}}(\hat{\beta}) - \text{Cov}(\hat{\beta})) \to 0 \tag{5.1}$$

We multiply by n since Slutsky's theorem requires sufficiently fast convergence of $\hat{V} \to V$. Since $\text{Cov}(\hat{\beta})$ itself shrinks at n^{-1}, we must have convergence of $\hat{V} \to V$ at a faster rate which we ensure by checking whether $n(\hat{V}-V) \to 0$.

Lemma 5.1

Assume D diagonal and $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,D)$. Let $\hat{\beta} = (X^TX)^{-1}X^Ty$. Estimate $\hat{D} = \text{diag}(\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2)$. Assume further that

1. $\frac{X^TX}{n} \to C$ as $n \to \infty$ where C is psd
2. $\sigma_i^2 < C_1 < \infty$ for all i, and $1/n \sum X_{ij}^2 \leq C_2 < \infty$ for all j

Then

$$\hat{\beta} \xrightarrow{p} \beta \text{ as } n \to \infty$$

Proof. We know that our estmated $\hat{\beta}$ is unbiased. We therefore only need to show that the covariance converges to 0.

We write

$$\text{Cov}\left(\hat{\beta}\right) = \frac{1}{n}\underbrace{\left(\frac{X^TX}{n}\right)^{-1}}_{(1)}\underbrace{\frac{X^TDX}{n}}_{(2)}\underbrace{\left(\frac{X^TX}{n}\right)^{-1}}_{(3)}.$$

We know that terms (1) and (3) are bounded by assumption 1. Further, term (2) is bounded by assumption 2 since any element of $\dfrac{X^TDX}{n}$ is weakly smaller than $C_1 \cdot C_2$. Therefore,

$$\text{Var}\left(\hat{\beta}_j\right) \to 0 \text{ as } n \to \infty, \ \forall j.$$

By Tchebshev's inequality,

$$\lim_{n\to\infty} P\left(\left|\hat{\beta}_j - \beta_j\right| > \varepsilon\right) = 0, \ \forall \varepsilon > 0$$

□

[18] Normality of $V^{-1/2}(\hat{\beta}-\beta)$ follows from our usual arguments of either the distribution of ϵ or a central limit theorem. The part of interest is now whether we can replace $V^{-1/2}$ with our estimated covariance matrix $\hat{V}^{-1/2}$.

Lemma 5.2

In addition, assume $\max_{i,j} |X_{ij}| \leq K < \infty$. Then

$$\frac{X^T \hat{D} X}{n} - \frac{X^T D X}{n} \xrightarrow{p} 0 \text{ as } n \to \infty$$

Proof. Consider an arbitrary elements (r, s). We can write

$$\frac{1}{n}\left(X^T \hat{D} X\right)_{rs} = \frac{1}{n} \sum_i X_{ir} X_{is} \hat{\varepsilon}_i^2 \tag{5.2}$$

Then,

$$\begin{aligned}
\frac{1}{n}\left(X^T \hat{D} X\right)_{rs} &= \frac{1}{n} \sum_i X_{ir} X_{is} \hat{\varepsilon}_i^2 \\
&= \frac{1}{n} \sum X_{ir} X_{is} \left(y_i - \hat{y}_i\right)^2 \\
&= \frac{1}{n} \sum X_{ir} X_{is} \left(\varepsilon_i - \underbrace{x_i^T (\hat{\beta} - \beta)}_{\to 0}\right)^2 \\
&= \frac{1}{n} \sum X_{ir} X_{is} \left(\varepsilon_i^2 + o_p(1)\right)
\end{aligned}$$

In the second last line we use our above result that $\hat{\beta} \to \beta$ in probability as $n \to \infty$.

To show consistency, we first show asymptotic unbiasedness.

$$\mathbb{E}\left[\frac{1}{n} \sum X_{ir} X_{is} \varepsilon_i^2\right] = \frac{1}{n} \sum X_{ir} X_{is} \sigma_i^2 = \frac{1}{n} (X^T D X)_{rs}$$

where we use (5.2) for population values (and in reverse).

Next, we show that the variance converges to 0 as $n \to \infty$,

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{n} \sum X_{ir} X_{is} \varepsilon_i^2\right) &= \frac{1}{n} \frac{1}{n} \sum_{i=1}^n X_{ir}^2 X_{is}^2 \operatorname{Var}\left(\varepsilon_i^2\right) \\
&= \frac{1}{n} \frac{1}{n} n \cdot \text{constant} \to 0 \ (n \to p)
\end{aligned}$$

For this, we may use that if $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ then $\operatorname{Var}\left(\varepsilon_i^2\right) = 3\sigma_i^4 < \infty$.

Thus, $\frac{X^T \hat{D} X}{n} - \frac{X^T D X}{n} \xrightarrow{p} 0$ as $n \to \infty$. □

This allows us to proof (5.1) since

$$n\left[\widehat{\operatorname{Cov}}\left(\hat{\beta}\right) - \operatorname{Cov}\left(\hat{\beta}\right)\right] = \left(\frac{X^T X}{n}\right)^{-1} \left[\frac{X^T \hat{D} X}{n} - \frac{X^T D X}{n}\right] \left(\frac{X^T X}{n}\right)^{-1}$$

where the mid term converges to zero and the other terms converge to some constant matrix C.

What is the risk that we "badly" estimate at least one element of the matrix if we also allow the number of covariates to grow? We can upper-bound this error by $p^2 \sup \Pr(\text{error}_i)$. Note that the probabiltiy of an error goes down at rate n^{-1}, as shown. If the number of covariates p grows at a smaller rate, e.g., $\log N$, this error probabiltiy still converges to 0.

One fundamental application of the HC estimator is when using linear models to approximate complex models. Even if we have $Y_i = f(X_i) + \varepsilon_i$ with ε_i independent of X_i, and $\text{Var}(\varepsilon_i) = \sigma^2$ when we use a linear approximation, we will have a heteroskedastic error.

In R: coeftest (fm, vcov = vcovHC(fm, type = "HC0") from sandwhich package.

5.3 Mixed Models

Another approach of modelling more complex error structures is to use mixed models. Suppose we have i individuals or clusters, and n_i observations per individual. Coefficients are likely to be different across people. One approach would be to have a block factor:

$$y_{ij} = (\beta_0 + \beta_{0,i}) + \beta_1 x_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N\left(0, \sigma^2\right) i.i.d$$

Here we only consider *fixed effects* and include a custom level-shift per cluster. With this we can easily make statements about individual intercepts but less easily derive statements about the underlying distribution.

A different approach uses so called random effects.

Random Intercept Model

$$\begin{aligned} &y_{ij} = (\beta_0 + u_i) + \beta_1 x_{ij} + \varepsilon_{ij}, \\ &\varepsilon_{ij} \sim N\left(0, \sigma^2\right), \; u_i \sim N\left(0, \sigma_1^2\right) \quad \text{i.i.d} \\ &\varepsilon_{ij} \perp u_k, \forall i, j, k \end{aligned}$$

I.e., we have a cluster-specific random shift (u_i).

This is a **mixed model** since we consider both fixed and random effects. We can easily study statements about the dispersion of of initial slopes in the population but not about indvidiuals.

It turns out that we implicitly model the correlation among samples of the same cluster.

Within block i:

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(u_i + \varepsilon_{ij}, u_k + \varepsilon_{ik}) \\ &= \text{Cov}(u_i, u_k) + \text{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) \\ &= \sigma_1^2 + \sigma^2 \mathbf{1}\{j = k\} \end{aligned}$$

Between blocks, we have zero covariance. Let $m \neq i$,

$$\text{Cov}(Y_{ij}, Y_{mk}) = \underbrace{\text{Cov}(\mu_i, \mu_m)}_{=0} + \underbrace{\text{Cov}(\varepsilon_{ij}, \varepsilon_{mk})}_{=0}$$

We therefore obtain a block-diagonal covariance matrix.

- Constant correlation structure
- Relative strength of within-subject correlation depends on relative magnitudes of σ^2 and σ_1^2

Random Intercept Random Slope Model

We can naturally extend this model to allow for variying slopes between clusters. We now not only consider a random shift in the intercept but also a random shift for the slope parameter.

$$
\begin{aligned}
&y_{ij} = (\beta_0 + u_{1,i}) + (\beta_1 + u_{2,i})\, x_{ij} + \varepsilon_{ij}, \\
&\varepsilon_{ij} \sim N\left(0, \sigma^2\right) i.i.d \\
&u_{1,i} \sim N\left(0, \sigma_1^2\right), \\
&u_{2,i} \sim N\left(0, \sigma_2^2\right), \\
&\operatorname{cor}\left(u_1, u_2\right) = \rho
\end{aligned}
$$

We initially estimate models using MLE to compare models. To obtain the final model we use restricted MLE (RMLE), as this removes the biased estimation of standard deviations.[19]

Note: The standard deviation of our slope and intercept estimates will approach the population values as $n \to \infty$, however, the standard error of the estimates will go to 0.

In R: lmer function in lme4 library.

Comparing Fixed Effects and Random Effects

Note that the question of fixed vs random effects depends fundamentaly on the question one wants to answer. If one is interested in the specific sample, fixed effects are more useful, while random effects allow inference about population attributes.

In terms of estimation, fixed effects are modelled by explicitly including a dummy variable for each cluster in the design matrix X and estimating coefficients for each cluster, e.g., α_i, throuhg OLS. On the other hand, to estimate random effects we use MLE and specify that the random effects follow a normal distribution from which we then find $\hat{\sigma}_1^2$.[20] Since we estimate fewer parameters, this allows scaling to larger number of random effects.

[19] We initially use MLE due to better properties of MLE and allowing comparison of non-nested models.

[20] Some libraries such as lme4 can still extract actual random effects by using a technique called *best linear unbiased prediction (BLUP)*. After estimating $\hat{\sigma}_1^2$, BLUPs are calculated for each random effect by taking conditional expectations of random effects given data and estimated coefficietns. The formula in a RI model is given by:

$$\hat{b}_j = \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2/n_j} \left(\bar{y}_j - \beta_0 - \beta_1 \bar{x}_j\right)$$

I.e., instead of estimating each coefficient directly, we estimate distribution parameters and then obtain random effects by taking conditional expectations.

General Linear Mixed Models

We can greatly increase complexity of such models by having more complicated random effects and more complex covariance structures. This gives rise to *General Linear Mixed Models.*

Consider y_i to be an n-dimensional vector for cluster i:

$$\begin{aligned} y_i &= x_i\beta + u_i\gamma_i + \varepsilon_i \\ \gamma_i &\sim \mathcal{N}(0, Q), \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\Sigma_{n_i}\right), \quad \gamma_i, \varepsilon_i \text{ indep.} \end{aligned}$$

We obtain the *conditional formulation* of our model by conditioning on random effects (i.e., we fix cluster i):

$$y_i \mid \gamma_i \sim \mathcal{N}\left(x_i\beta + u_i\gamma_i, \sigma^2\Sigma_{n_i}\right)$$

The *marginal formulation* subsums random effects and other noise terms:

$$\begin{aligned} &y_i = X_i\beta + \varepsilon_i^*; \quad E\left(y_i\right) = X_i\beta \\ &\operatorname{Cov}\left(y_i\right) = \operatorname{Cov}\left(u_i\gamma_i\right) + \operatorname{Cov}\left(\varepsilon_i\right) = u_iQu_i^\top + \sigma^2\Sigma_{n_i} \\ &\Rightarrow Y_i \sim \mathcal{N}\left(x_i\beta, \sigma^2\Sigma_{n_i} + u_iQu_i^\top\right) \end{aligned}$$

Example 5.1

To recover the RI model from the above, use

$$u_i = (1, \ldots, 1)^\top,\ Q = \sigma_1^2,\ \Sigma = \mathbb{I}$$

We would then recover the exact same structure we had above.

Residual analysis is more difficult, as the complicated error structure leads to many technicalities. We can investigate Turkey-Anscombe plots but should only consider major deviations. We can also use QQ-plots on the estimated random effects to investigate whether the normality assumption holds.

To summarise: A general linear mixed model is given by

$$y = X\beta + U\gamma + \varepsilon$$

with

$$\begin{pmatrix}\gamma \\ \varepsilon\end{pmatrix} \sim N\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}\right).$$

In this model, $\mathbf{X}$ and $\mathbf{U}$ are design matrices, β is a vector of fixed effects, and γ is a vector of random effects. The covariance matrices for γ and ε are assumed to be nonsingular, and therefore positive definite, and γ and ε are independent.

$$\mathbf{U} = \operatorname{blockdiag}\left(\mathbf{U}_1, \ldots, \mathbf{U}_i, \ldots, \mathbf{U}_m\right) = \begin{bmatrix} \mathbf{U}_1 & & & \mathbf{0} \\ & \ddots & & \\ & & \mathbf{U}_i & \\ & & & \ddots \\ \mathbf{0} & & & \mathbf{U}_m \end{bmatrix}$$

$\varepsilon \sim N(\mathbf{0}, \mathbf{R}), \gamma \sim N(\mathbf{0}, \mathbf{G})$ with block diagonal covariance matrices:

$$\mathbf{R} = \text{blockdiag}\left(\sigma^2\mathbf{\Sigma}_{n_1}, \ldots, \sigma^2\mathbf{\Sigma}_{n_i}, \ldots, \sigma^2\mathbf{\Sigma}_{n_m}\right)$$
$$\mathbf{G} = \text{blockdiag}(\mathbf{Q}, \ldots, \mathbf{Q}, \ldots, \mathbf{Q})$$

Note: dimensions of above elements

- $\boldsymbol{y}_i$ is an $n_i \times 1$ response vector.
- $\boldsymbol{X}_i$ is an $n_i \times p$ fixed-effects design matrix.
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed-effect coefficients (common across groups).
- $\boldsymbol{U}_i$ is an $n_i \times q_i$ random-effects design matrix. E.g., this could consist of one column of ones and one column with time observations in a typical RIRS example.
- $\boldsymbol{\gamma}_i$ is a $q_i \times 1$ vector of random-effect coefficients. E.g., we assume that there is some randomness in initial reaction time and response to days of sleep deprivation between people. $\boldsymbol{\gamma}_i$ is the "drawn" value from this random distribution which specifies initial reaction and reaction to sleep deprivation for individual i.
- ε_i is an $n_i \times 1$ vector of random errors.

5.4 Comparison between different approaches for non iid errors

- Generalised Least Squares:
  - applicable to heteroskedastic and correlated erros but requires knowledge of covariance matrix up to multiplicative factor.
  - Different coefficient estimates.
- Two-stage or MLE:
  - applicable to heteroskedastic and correlated erros but requires a parametric form of covariance structure (e.g., AR process) (do not require exact covariance matrix ex ante).
  - Different coefficient estimates.
- Huber-White:
  - Applicable to heteroskedastic but uncorrelated errors.
  - Note: same coefficient estimates as OLS
  - Heteroskedasticity and autocorrelation robust estimators exist.
- Mixed-models:
  - Implicitly models error structure through random effect structure, can include both heteroskedastic and correlated errors but requires knowledge of grouping of data.
  - Different coefficient estimates.
  - Higher variance compared to other approaches but less detailed knowledge about covariance matrix required.

6 Generalized Linear Models

Generalized Linear Models (GLMs) model relationships between explanatory variables and the parameter of the distribution. That is, we assume our target variable follows some distribution $F(\theta(X))$ whose parameters depend on X.

$$\begin{aligned} S: &\quad Y \sim F(\theta(X)) \\ D: &\quad g(\theta(X)) = \beta_0 + \beta_1 x \end{aligned}$$

Note that while Y is stochastic, we view the link between the parameter and our covariates as being deterministic (no error term).[21]

Three key components:

- Link function $g(\cdot)$
- Linear (in β) predictor
- Distribution of Y

We will see that GLM usually assumes that

- $F(\theta)$ is in exponential family
- θ is expected value
- Every $F(\theta)$ then has a canonical link function with 'nice' properties

Example 6.1 (Linear Regression as GLM)

Linear regression is a special case of GLM where we model Y to be normally distributed, with the mean depending on X.

$$\begin{aligned} &\mathrm{S}: Y \sim N\left(\mu(x), \sigma^2\right) \\ &\mathrm{D}: \mu(x) = \beta_0 + \beta_1 x \end{aligned}$$

Example 6.2 (Logistic Regression as GLM)

Assume we have a binary outcome variable. We then naturally model this as a binomial distribution:

$$\mathrm{S}: Y \sim \mathrm{Bin}(1, p(x))$$

We further assume that p depends on explanatory variables X. We may model the relationship between X and p as:

$$p(x) = \frac{\exp\left(\beta_0 + \beta_1 x\right)}{1 + \exp\left(\beta_0 + \beta_1 x\right)}$$

[21] I.e., we do not explicitly include or model an error term ε. Previously, randomness in $Y \mid X$ only depended on the error, we now instead model Y to follow the distribution $F(\theta(X))$ directly.

where the right hand term is the *Logistic function.* Then

$$\mathrm{D} : \log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x.$$

We call D the *link function.*
We could use any *cdf* since this maps $X \to [0,1]$ Logistic regression is popular since it models *Log-Odds.*
Recall that odds of an event A are $P(A)/(1-P(A))$. We note that log-odds and odds grow monotonically with probability. Log-odds are directly modelled in Logistic regression.
A related concept is the risk ratio. If we are intersted in some event A given B, the risk ratio is $P(A|B)/P(A|B^C)$. The odds of $A|B$ are $P(A|B)/P(A^C|B) = P(A|B)/\left[1-P(A|B)\right]$.

6.1 Logistic Regression

We now study logistic regression in more detail. Note that other than Logistic regression, any CDF can be used to model the link between p and predictors X as it ensures that the predicted probability is in the interval $[0,1]$.

Logistic regression as latent variable model

Logistic regression can be thought of as a latent variable model. Assume $Z_i = x_i^T\beta + \varepsilon_i$, where ε_i i.i.d. and symmetric around 0. But we only observe

$$\begin{aligned}
&\to Y_i = 1, \text{ if } Z_i > 0 \\
&\to Y_i = 0, \text{ if } Z_i \leq 0
\end{aligned}$$

We can then see

$$\begin{aligned}
P(Y_i = 1) &= P(x_i^T\beta + \varepsilon_i > 0) \\
&= P(\varepsilon_i > -x_i^T\beta) \\
&= P(\varepsilon_i < x_i^T\beta) (\text{by symmetry})
\end{aligned}$$

If we assume $\varepsilon_i \sim \text{Logistic}(\mu = 0, s = 1)$,[22] then

$$\begin{aligned}
&\underbrace{P(Y_i = 1)}_{=:\pi} = P(\varepsilon_i < x_i^T\beta) = \frac{\exp\{x_i^T\beta\}}{1+\exp\{x_i^T\beta\}} \\
\implies &\log\left(\frac{\pi}{1-\pi}\right) = x_i^T\beta
\end{aligned}$$

[22] Note that the Logistic distribution has pdf $f(x;\mu,s) = \frac{e^{-(x-\mu)/s}}{s\left(1+e^{-(x-\mu)/s}\right)^2}$, mean is given by μ, standard deviation $s^2\pi^2/3$. s is called the scale parameter. The CDF for the standard logistic distribution is $P(\varepsilon_i \leq x) = F(x) = \frac{e^x}{1+e^x}$.

We can instead also assume that $\varepsilon_i \sim \mathcal{N}(0,1)$ and arrive at **Probit regression.**[23]

If we have more than two *ordered* categories, we can use proportional odds logistic regression.

Parameter Estimation

We use maximum likelihood estimation, given a vector x_i of explanatory variables and binary outcome y_i. I.e., we find $\hat{\beta}$ to maximize

$$l(\hat{\beta}) = \prod_{i:y_i=1} p(x_i) \prod_{j:y_j=0} (1 - p(x_j)) = \prod_{i:y_i=1} \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T \beta)} \prod_{j:y_j=0} \frac{1}{1+\exp(x_i^T \beta)}.$$

We now assume just one regressor and more succinctly write this as

$$L(y_1, \ldots, y_n; \beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{(1-y_i)}$$

with $\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$ and $\eta_i = x_i^T \beta$. Taking logs:[24]

$$\begin{aligned} \ell(y_1, \ldots, y_n; \beta) &= \sum y_i \log \pi_i + (1-y_i) \log(1-\pi_i) \\ &= \sum_i \left(y_i x_i^\top \beta - \log\left(1 + e^{x_i^\top \beta}\right)\right) \end{aligned}$$

Consider the partial derivatives,

$$\frac{\partial \pi_i}{\partial \beta_0} = \frac{\partial \pi_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_0} = \pi_i(1-\pi_i); \; \frac{\partial \pi_i}{\partial \beta_1} = x_i \pi_i (1-\pi_i)$$

Further, let $n\ell = -\ell$. We aim to minimize the negative log-likelihood,

$$\begin{aligned} \frac{\partial n\ell}{\partial \beta_0} &= -\sum y_i \frac{1}{\pi_i} \pi_i(1-\pi_i) + (1-y_i) \frac{(-1)}{1-\pi_i} \pi_i (1-\vec{\pi}_i) \\ &= -\sum y_i (1-\pi_i) - (1-y_i)\pi_i = -\sum y_i - \pi_i \\ \frac{\partial n\ell}{\partial \beta_1} &= -\Sigma (y_i - \pi_i) x_i \end{aligned}$$

Thus,

$$\nabla_\beta n\ell = \begin{pmatrix} \partial n\ell/\partial\beta_0 \\ \partial n\ell/\partial\beta_1 \end{pmatrix} = \begin{pmatrix} \sum(\pi_i - y_i) \\ \sum x_i(\pi_i - y_i) \end{pmatrix}$$

We will not be able to set these equations equal to zero and solve them analytically. We therefore consider Newton-Raphson instead.

[23] The link function affects units of estimated parameters. Due to different variances, estimated parameters for Probit and Logit regression will be different. If we choose scale parameters such that $\mathrm{Var}(\epsilon_i)$ is identical for Logit and Probit, then we will obtain very similar results. Also, no matter the scale, estimated probabilities will be similar.

[24] Use $\ln(1-\pi_i) = \ln\left(1 - \frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) = \ln\left(\frac{1}{1+\exp(\eta_i)}\right)$.

Second-order derivatives:

$$\frac{\partial^2 nl}{\partial \beta_0^2} = \frac{\partial}{\partial \beta_0}\left(\sum (\pi_i - y_i)\right) = \sum \frac{\partial}{\partial \beta_0}\pi_i = \sum \pi_i (1-\pi_i)$$
$$\frac{\partial^2 nl}{\partial \beta_1^2} = \ldots = \sum x_i^2 \pi_i (1-\pi_i)$$
$$\frac{\partial^2 nl}{\partial \beta_1 \partial \beta_1} = \ldots = \sum x_i \pi_i (1-\pi_i)$$

The Hessian matrix is therefore given by

$$H_{n\ell} = \begin{bmatrix} \sum \pi_i (1-\pi_i) & \sum x_i \pi_i (1-\pi_i) \\ \sum x_i \pi_i (1-\pi_i) & \sum x_i^2 \pi_i (1-\pi_i) \end{bmatrix} = \sum_{i=1}^{N} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \pi_i (1-\pi_i) \begin{bmatrix} 1 & x_i \end{bmatrix} \tag{6.1}$$

We can use Newton's method to estimate $\hat{\beta}$, using some initial value:

$$\beta_{n+1} = \beta_n - H_{n\ell}^{-1} \nabla_{n\ell}$$

Note that y does not appear in Eq. (6.1). Therefore, taking expectation is identical to (6.1) itself. Recalling the definition of Fisher information (Definition A.1 and Eq. (A.1)), the Fisher information matrix is given by

$$J_{ij}(\beta) = -\mathbb{E}\left(\frac{\partial^2}{\partial \beta_i \partial \beta_j} \ell\left(y_1, \ldots, y_n, \beta\right)\right)$$

Thus, Newton-Method is identical to Fisher scoring. This is always the case when canonical links are used. Futher, the optimisation problem is convex and therefore has a global minimum.

Inference

Due to MLE estimation, we have asymptotic normality. In particular,

$$\widehat{\theta} \overset{\text{a}}{\sim} \mathcal{N}(\theta, V(\theta)),$$

where the asymptotic covariance matrix $V(\theta)$ of $\widehat{\theta}$ is the inverse of the Fisher information which we calculated for the simple case of one predictor in (6.1). The general form is given by

$$V(\theta)^{-1} = I(\theta) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \mathbb{E}\left[\left(y_i - P_\theta\left[Y_i = 1\right]\right)^2\right] = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \frac{\exp\left(\mathbf{x}_i^T \theta\right)}{\left(1 + \exp\left(\mathbf{x}_i^T \theta\right)\right)^2}.$$

We can test nested models based on the *Likelihood Ratio test.* If we consider two nested models with $q < p$ dimension, then

$$2\left(\ell\left(\widehat{\theta}^{(p)}\right) - \ell\left(\widehat{\theta}^{(q)}\right)\right)$$

is known to be asymptotically χ^2 with $(p-q)$ degrees of freedom.

In particular, it is often useful to compare our model to the worst model with only an intercept, $\ell\left(\hat{\beta}^0\right)$ and the best possible models with as many parameters as observations, $\ell\left(\hat{\beta}^S\right)$ (*saturated model*).

- **Null deviance:** $\mathbf{D}_0 = 2\left(\ell\left(\hat{\beta}^S\right) - \ell\left(\hat{\beta}^0\right)\right)$
- **Residual deviance:**

$$2\left(\ell\left(\hat{\beta}^S\right) - \ell(\hat{\beta})\right) = \cdots = \sum 2\left(y_i \log\left(\frac{\hat{\pi}_i^S}{\hat{\pi}_i}\right) + (1-y_i)\log\left(\frac{1-\hat{\pi}_i^S}{1-\hat{\pi}_i}\right)\right)$$

Note that we can evaluate the contribution of each observation to the residual deviance. We can use this to analyse residuals which is much more tricky than in linear regression.

We can compare nested models using AIC.

Interpretation

The estimated β_i coefficient can be interpreted as follows.

- **Odds-scale**: If x_i increases by one unit, *odds* increase by **factor** $\exp\{\beta_i\}$.
- **Log-odds scale:** If x_i increases by one unit, *log-odds* increase by β_i.
- Note: it is difficult to make simple and compact statement in terms of the probability scale.

In R:

- fm1 <− glm(y ~ x, data = df, family = binomial())
- Predict response either in terms of log-odds or on response scale, by specifying predict.glm(..., type ="response") or type ="link"

6.2 Poisson Regression

Suppose our Y_i is count data. In this case we may assume $Y_i \sim Poisson(\lambda_i)$. Recall that both variance and expectation of Y_i are given by λ_i.

Poission regression assumes that the parameter λ_i can be explained by certain variables. We model

$$\begin{aligned} &\text{D: } \log(\mu_i) = \log(\lambda_i) = x_i^T\beta \\ &\text{S: } Y_i \sim \text{Pois}(\lambda_i) \end{aligned}$$

Note that we use the log(.) function to ensure non-negativity of our estimated λ_i.

Estimation

We estimate our model using MLE.

Log-likelihood:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i} \text{ where } \lambda_i = \exp\{\beta_0 + \beta_1 x_i\}$$

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \log \lambda_i - \log(y_i!) - \lambda_i =$$

$$= \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \log(y_i!) - e^{\beta_0 + \beta_1 x_i}$$

Score equations:

$$s\left(\hat{\beta}_0, \hat{\beta}_1\right) = \begin{pmatrix} \frac{\partial l}{\partial \beta_0} \\ \frac{\partial l}{\partial \beta_1} \end{pmatrix} \stackrel{!}{=} 0 \Rightarrow \hat{\beta}_0, \hat{\beta}_1$$

where

$$\frac{\partial \ell}{\partial \beta_0} = \sum y_i - \underbrace{e^{\beta_0 + \beta_1 x_i}}_{\lambda_i} \stackrel{!}{=} 0$$

$$\frac{\partial l}{\partial \beta_1} = \sum y_i x_i - x_i e^{\beta_0 + \beta_1 x_i} = \sum x_i (y_i - e^{\beta_0 + \beta_1 x_i}) \stackrel{!}{=} 0$$

Note that this is similar in the form to normal equations in linear regression where $X^T(y - \hat{y}) = 0$. This system of equations is non-linear, we use numerical optimization (Fisher Scoring). We will later show that this is a convex optimization problem.

Hessian matrix

$$\frac{\partial^2 l}{\partial \beta_0^2} = -\sum \lambda_i; \; \frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = -\sum x_i \lambda_i; \; \frac{\partial^2 l}{\partial \beta_1^2} = -\sum x_i^2 \lambda_i$$

Therefore,

$$F(\beta_0, \beta_1) = -\sum \lambda_i \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} = -\sum \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} \lambda_i$$

Fisher Scoring. Note that our hessian matrix does not depend on y. Thus taking expectation is identical to directly taking the matrix. Let $\mathcal{I}(\beta_0, \beta_1) = -F(\beta_0, \beta_1)$ be the Fisher matrix, we have

$$\hat{\beta}^{(t+1)} = \hat{\beta}^{(t)} + \mathcal{I}(\hat{\beta}^{(t)})^{-1} s(\hat{\beta}^{(t)})$$

This optimization process can be rewritten as iteratively re-weighted least squares. This holds true for all GLMs (i.e., all exponential distributions) when using the canonical link.

Inference

Using asymptotic properties of the MLE, we have $\hat{\beta} \overset{a}{\sim} \mathcal{N}\left(\beta, \mathcal{I}^{-1}(\hat{\beta})\right)$.

Model Comparison

- Null model: worst model with only intercept

$$\Rightarrow l\left(\hat{\beta}_0\right) = \Sigma y_i \hat{\beta}_0 - \log\left(y!\right) - e^{\hat{\beta}_0}$$

- Saturated model (best"): #pars = #obs

$$\Rightarrow y_i = \lambda_i$$
$$\Rightarrow l\left(\hat{\beta}^s\right) = \begin{cases} \sum y_i \log\left(y_i\right) - \log\left(y_i!\right) - y_i & \text{if } y_i \neq 0 \\ 0 & \text{else} \end{cases}$$

We then calculate

- Null deviance: $-2\left(l\left(\hat{\beta}^s\right) - \ell\left(\hat{\beta}^0\right)\right)$
- Residual deviance: $-2\left(l\left(\hat{\beta}^s\right) - l(\hat{\beta})\right)$

We can write the residual deviance as a sum of **squared deviance residuals** of each observation.

Caveat: Poission assumes linear relation between expectation and variance with slope 1. One may use quasi-poisson to loosen the realtionship to a linear relationship with slope $\neq 1$. Using quassi-poisson leads to same point estimates but different estimates for standard errors.[25] Large overdispersion should lead to use of other models, e.g., negative binomial regression.

6.3 Gamma Regression

Recall that if we have positive continuous data we may want to log-transform the data. The issue in interpreting effect size is that

$$g^{-1}(E(g(Y))) \neq E(Y)$$

However, if we have a *symmetric* distribution, as log(.) preserves ordering, the expectation equals median so we can make statements about changes in median.

GLM offers an alternative (for the case of skewed distributions) by considering gamma regressions. For each level of the explanatory variable, we find μ_i based on x_i and model a separate gamma distribution with term μ_i.

- Gamma distribution $\Gamma(k, \theta)$ where k measures shape and θ scale

$$E(X) = k\theta, \text{Var}(X) = k\theta^2$$

- Gamma regression then models

$$\text{D} : \log\left(\mu_i\right) = x_i^T \beta$$
$$\text{S} : Y_i \sim \Gamma\left(\mu_i, v\right)$$

[25] Since we estiamte the quasi-dispersion parameter, we will obtain t instead of z values.

6.4 Overview: Generalized Linear Models

All previous cases are examples of Generalized Linear Models which we now summarize.

- Three components:
  - Distribution from exponential family
  - link function (preferably use canonical link)
  - linear predictor
- D : $\mathrm{g}(\mu(x)) = x_i^T\beta$ (or $\mu(x) = h\left(x_i^T\beta\right)$, thus $g = h^{-1}$)
- S : $Y \sim F(\mu)$

Recall: Exponential family is given by distributions with density of the form

$$f(y \mid \theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)\right)$$

The log-density is given by

$$\log f(y \mid \theta) = \frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w).$$
$$\mathbb{E}(y) = \mu = b'(\theta), \quad \mathrm{Var}(y) = \phi b''(\theta)/w.$$

here θ is related to the linear predictor, ϕ is a dispersion parameter, and c a nuisance parameter.

Example 6.3 (Bernoulli is in exp. family)
We have $P(X = x) = \pi^x(1-\pi)^{(1-x)}$. We can rewrite this by taking log and exponential:

$$\exp\left\{x \log \frac{\pi}{1-\pi} + \log(1-\pi)\right\}$$

we can then set $\theta = \log \frac{\pi}{1-\pi}$, $-b(\theta) = \log(1-\pi)$ and $w = \phi = c = 1$.

MLE is done in the same spirit as above, by finding roots of score functions using Fisher scoring. Note: Fisher scoring can be written as **iteratively re-weighted least squares (IRLS)** estimates whereby the weights depend on the derivative of the link function, the variance function of the distribution, and the current estimate.

One can easily show that for exponential family densities, the log likelihood is concave, so one can find a global optimum:

$$l(\theta) = \frac{y\theta - b(\theta)}{\phi}\omega + c(y, \phi, \omega)$$
$$l'(\theta) = \frac{\partial l}{\partial \theta} = \frac{y - b'(\theta)}{\phi}\omega$$
$$l''(\theta) = \frac{\partial^2 l}{\partial \theta^2} = -\frac{b''(\theta)}{\phi}\omega = -\frac{\mathrm{Var}(Y)\omega^2}{\phi^2} \leq 0$$

We can also extend GLMs to include mixed effects which leads to Generalied Linear Mixed Models.

7 Extensions

7.1 Non-Linear Regression

Consider

$$y_i = f(x_i, \theta) + \varepsilon_i,$$

- p: Number of parameters (θ),
- m: Number of explanatory variables (x)

where potentially $m \neq p$.

Assumptions on residual term: Usually one of the two below

- $\varepsilon_i \sim N(0, \sigma^2)$ iid (then OLS = MLE $\rightarrow$ properties of MLEs), or
- $E(\varepsilon_i) = 0$ and $\text{Cov}(\varepsilon_i) = \sigma^2 I \rightarrow$ asymptotic normality of pars under technical assumptions

Usually the form of $f(x, \theta)$ is given by *context*. That is, we assume to know $f(\cdot)$.

7.1.1 Linearizing

In many cases, we may be able to linearize a non-linear function and thus obtain coefficients and predictions in our usual OLS framework. Example:

Michaelis-Menten $y = \frac{\theta_1 x}{\theta_2 + x} \rightarrow \frac{1}{y} = \frac{\theta_2}{\theta_1} \cdot \frac{1}{x} + \frac{1}{\theta_1}$ can be linearized to $\tilde{y} = \frac{1}{y}, \tilde{x} = \frac{1}{x} \rightarrow \tilde{y} = \frac{\theta_2}{\theta_1} \cdot \tilde{x} + \frac{1}{\theta_1}$

- prediction is easily done
- however, mainly interested in parameters:
  - $\rightarrow \theta_1$ easy to get from $\frac{1}{\theta_1}$; θ_2 much harder to untangle from θ_2/θ_1
- Error structure might not be additive anymore
  - $\text{y} = \frac{\theta_1 x}{\theta_2 + x} + \varepsilon \rightarrow \tilde{y} = \frac{\theta_2}{\theta_1} \cdot \tilde{x} + \frac{1}{\theta_1} + \, ?$

Instead use GLM with inverse link (if error structure is additive):

$$\frac{1}{\mu} = \frac{\theta_2}{\theta_1} \cdot \frac{1}{x} + \frac{1}{\theta_1} \text{ and } Y \sim N\left(\mu(x), \sigma^2\right)$$

Error structure is important as violation leads to wrong coverage probabilities of CI!

We can investigate deviations from assumed error structure by investigating the Turkey-Anscombe plot. If we find deviations, we may want to instead consider fitting a nonlinear regression model.

7.1.2 Model fitting: Iterative Least Squares

If we have n observations, we generate a vector $\eta(\theta)$ which contains $f(x,\theta)_i$ for each observation i. Our vector of predictions $\eta(\theta)$ and our true y vector are (each) a point in $\mathbb{R}^n$. As we change θ, we generate a new vector and we trace out a model surface in $\mathbb{R}^n$. We then want to find the point on the model surface closest to actual observations. I.e., we want to find $\hat{\theta}$ to minimize

$$S(\theta) = \sum_{i=1}^{n} (y_i - \eta_i(\theta))^2 . \tag{7.1}$$

We need to use *numerical optmization* as this generally does not allow for a closed-form solution.

Below we denote by $A(\theta)$ the Jacobian Matrix of $\eta(\theta)$.[26]

Algorithm (Gauss-Newton Method)

1. Start with initial value $\hat{\theta}^{(0)}$
2. For $k = 1, 2, \ldots$
   - Calculate tangent plane of $\eta(\theta)$ at $\hat{\theta}^{(k-1)}$:
   $$\eta(\theta) \approx \eta\left(\hat{\theta}^{(k-1)}\right) + A\left(\hat{\theta}^{(k-1)}\right) \cdot \left(\theta - \hat{\theta}^{(k-1)}\right)$$
   - Project on tangent plane (linear regression problem):
   Note: $Y = n(\theta) + \varepsilon$, thus
   $$Y = \eta(\theta) + \varepsilon \approx \eta\left(\hat{\theta}^{(k-1)}\right) + A\left(\hat{\theta}^{(k-1)}\right) \cdot \left(\theta - \hat{\theta}^{(k-1)}\right) + \varepsilon$$
   $$Y - \eta(\hat{\theta}^{(k-1)}) \approx A\left(\hat{\theta}^{(k-1)}\right) \cdot \left(\theta - \hat{\theta}^{(k-1)}\right) + \varepsilon$$
   «Preliminary residuals»: $\tilde{Y}^{(k-1)} = Y - \eta\left(\hat{\theta}^{(k-1)}\right)$
   «correction term»: $\beta^{(k-1)} = \left(\theta - \hat{\theta}^{(k-1)}\right)$
   Solve:
   $$\tilde{Y}^{(k-1)} = A\left(\hat{\theta}^{(k-1)}\right) \cdot \beta^{(k-1)} + \varepsilon$$
   This is a linear regression which yields $\beta^{(k-1)}$.
   - Use $\beta^{(k-1)}$ to update estimated coefficients:
   $$\hat{\theta}^{(k)} = \hat{\theta}^{(k-1)} + \beta^{(k-1)}$$
3. Iterate until convergence

That is, at each iteration we're using a linear approximation by a tangent plane. We then solve a "linear regression problem" with the preliminary residuals as our target value and the Jacobian matrix at our parameter estimate as regressors. We use the estimated coefficients $\hat{\beta}$ as our correction term to update our coefficients $\hat{\theta}$.

Convergence depends on starting values since model surface might be arbitratily complex (thus non-convex, local optima). Three methods for choosing starting values:

[26] I.e., the i-th column of A is the gradient $d\eta(\theta)/d\theta_i$. If $\eta(\theta) = X\beta$, then $A(\theta) = X$ so we recover the usual OLS problem in the below algorithm.

1. From context or experience
2. From linearized functions (OLS fit)
3. From data (parameters have meaning.)

Self-starting functions automate this to find initial values.

Example 7.1 (Interpreting coefs in Biochemical Oxygen Demand (BOD))

Model:

$$f(x, \theta) = \theta_1 \left(1 - \exp\left(-\theta_2 x\right)\right); \text{ assume } \theta_1 > 0, \theta_2 > 0$$

Interpretation of parameters:

$$- \lim_{x \to \infty} f(x, \theta) = \theta_1 \to \text{ Asymptote}$$
$$- \frac{df}{dx}(x = 0) = \theta_1 \theta_2 \to \text{ Slope at } x = 0$$

In R: fm2 <− nls(yObs ~ SSgompertz(x, Asym, b2, b3), data = df).

Manually:

fm <− nls(yObs ~ t1∗ exp(−t2∗ t3^x), data = df, start = c(t1 = 12, t2 = 5, t3 = 0.5))

7.1.3 Inference

In general, inference and assessing model fit is more difficult in nonlinear least squares.[27] Nevertheless, there are multiple useful approaches for inference in nonlinear problems.

Based on Linear Approximation

- From Gauss-Newton, we approximate the model surface by the tangential plane given by $A(\hat{\theta})$
- Fit standard linear regression: $Y = A(\hat{\theta}) \cdot \beta + \varepsilon$
- One can show that at optimum with $\hat{\beta} \approx 0$, $\text{Cov}(\hat{\theta}) \approx \text{Cov}(\hat{\beta}) = \sigma^2 \left(A\left(\theta_0\right)^T A\left(\theta_0\right)\right)^{-1}$. Therefore,
$$\widehat{\theta} \overset{\text{asym.}}{\sim} \mathcal{N}\left(\theta_0, \sigma^2 \left(A\left(\theta_0\right)^T A\left(\theta_0\right)\right)^{-1}\right)$$
where we in practice replace θ_0 with $\hat{\theta}$.
- Thus we get for a $(1 - \alpha)$-confidence interval:
$$\widehat{\theta}_k \pm t_{n-p;1-\alpha/2} \operatorname{se}\left(\widehat{\theta}_k\right)$$
where $\operatorname{se}\left(\widehat{\theta}_k\right) = \widehat{\sigma} \sqrt{\left(\left(A(\widehat{\theta})^T A(\widehat{\theta})\right)^{-1}\right)_{kk}}$ and $\hat{\sigma}^2 = \frac{s(\widehat{\theta})}{n-p} = \frac{1}{n-p} \sum_{i=1}^{n} \left(y_i - \eta_i(\widehat{\theta})\right)^2$

[27] E.g., is the assumption for Gompertz growth really satisfied? → hard to answer from fit. But sometimes there are nested models which we can test.

- Our test statistic is:

$$\delta_k(\theta_k^*) = \frac{\hat{\theta}_k - \theta_k^*}{\hat{\text{s.e.}}(\hat{\theta}_k)} \approx t_{n-p} \tag{7.2}$$

Note: CI symmetric by design

Based on Likelihood

Idea: Points where drop of in likelihood from optimum is small are still very likely. Use dropoff in likelihood ("vertical distance") to find confidence set.[28]

- Assume $H_0 : \theta = \theta^*$ is true which imposes q restrictions[29]
- Then it approximately holds:

$$T = \frac{n-p}{q} \cdot \frac{S(\theta^*) - S(\hat{\theta})}{S(\hat{\theta})} \approx F_{q,n-p}$$

- This is still only an approximate result; but in practice often much more accurate than linear approximation
- Construct confidence regions: Collect all vectors θ^* that are not rejected by the F -Test $\rightarrow$ difficult to visualize and summarize.

Alternative, to obtain tests for single parameters, we can use **profiling**

- Assume $H_0 : \theta_k = \theta_k^*$
- Fix parameter of interest at some arbitraty value $\theta_k = \theta_k^*$ and minimize $S(\theta)$ wrt. $\theta_j, j \neq k$; denote the minimum by $\tilde{S}_k\left(\theta_k^*\right)$
  - Repeating for all possible values of θ_k, then the function $\tilde{S}_k\left(\theta_k\right)$ is the "*profile likelihood*"
- Under $H_0 : \theta_k = \theta_k^*$, it approximately holds: $\tilde{T}_k\left(\theta_k^*\right) = (n-p) \cdot \frac{\tilde{S}_k\left(\theta_k^*\right) - S(\widehat{\theta})}{S(\widehat{\theta})} \approx F_{1,n-p}$

[28] Note that $S(\theta)$ is not a likelihood. We can, however, recover the likelihood and create a link between the two values: Under the assumption that errors are normally distributed $\left(\varepsilon_i \sim N\left(0, \sigma^2\right)\right)$, the likelihood function is:

$$L\left(\theta, \sigma^2\right) = \prod \left(1/\sqrt{2\pi\sigma^2}\right) \exp\left(-\left(y_i - f\left(x_i, \theta\right)\right)^2 / \left(2\sigma^2\right)\right)$$

Taking the log of this gives us the log-likelihood:

$$\ell\left(\theta, \sigma^2\right) = \text{ constant } - (n/2) \log\left(\sigma^2\right) - \left(1/\left(2\sigma^2\right)\right) \sum \left(y_i - f\left(x_i, \theta\right)\right)^2$$

Notice that the last term contains $S(\theta) = \sum \left(y_i - f\left(x_i, \theta\right)\right)^2$, our sum of squared residuals. (For $n \rightarrow \infty$ the F - test is the same as the likelihood-ratio test, and the sum of squares is, up to a constant, equal to the negative log-likelihood).

[29] Note that we have some inconsistent notation. In linear regression, we imposed $p - q$ restrictions in our submodel. Here we instead denote by q itself the number of restrictions in the submodel.

- Due to relation between $F_{1,n-p}$ and t_{n-p} this corresponds to:

$$T_k\left(\theta_k^*\right) = \operatorname{sign}\left(\hat{\theta}_k - \theta_k^*\right) \cdot \frac{\sqrt{\tilde{S}_k\left(\theta_k^*\right) - S(\widehat{\theta})}}{\widehat{\sigma}} \approx t_{n-p} \tag{7.3}$$

- CI:

$$\left\{\theta_k^* \mid \sqrt{\tilde{S}_k\left(\theta_k^*\right) - S(\widehat{\theta})} \leq t_{n-p;1-\alpha/2}\widehat{\sigma}\right\}.$$

- CI based on profiling are usually not symmetric[30] and perform better than CI based on linear approximation
- We can apply a montone transformation to θ_k and apply the same monotone transformation to the CI. This does not hold for the linear approximation.

7.1.4 Assessing non-linearity

We can test how well the linear approximation works. To do so, we compare the two tests based on linear approximation and profiling.

- If linear approximation is good, $\delta_k\left(\theta_k^*\right)$ and $T_k\left(\theta_k^*\right)$ should behave similarly
- Check in profile t-plot: Plot $T_k\left(\theta_k^*\right)$ against $\delta_k\left(\theta_k^*\right)$ (or $T_k\left(\theta_k^*\right)$ against θ_k^*) when varying θ_k^*
- In a linear setting, we would expect a diagonal straight line if $T_k\left(\theta_k^*\right)$ against θ_k^*
- In a strongly non-linear setting, we should see a clear deviation from this

Additionally, we can investigate likelihood profile traces. I.e., we fix arbitrary values for θ_1 and plot the optimal θ_2 (and vice versa). So for each point on the line, we're seeing what value of θ_2 gives us the best model fit when we force θ_1 to take a specific value. This creates a "trace" showing how the optimal values of the parameters move together.

When the lines in a profile trace are nearly parallel, it suggests that the parameters are highly correlated or potentially redundant in the model.

7.2 Non-parametric regression

We are in a scenario where

$$Y = f(x) + \epsilon$$

with $\mathbb{E}(\epsilon) = 0, \operatorname{Cov}(\epsilon) = \sigma^2\mathbf{1}$ and f twice continuously differentiable.

Difference to non-linear regression: We do not know the form of the nonlinear function f. Often in this case polynomial regressions are unstable and nonparametric approaches produce better results.

[30] Note that compared to the linear test statistic in (7.2), here $s(\theta)$ enters the test, not θ directly. I.e., we compare the "vertical distance" of the likelihood rather than the "horizontal distance" between parameter estimates.

7.2.1 Kernel Regression

- For each x, we average the observed y-values in neighborhood (e.g. $x \pm h$) where h is a fixed bandwidth
- We weight observations inversely to the distance from x (i.e., lower weights on more distants obs)
- Define weight via "Kernel function" (often a probability density function)
  - Problem:
    $$\hat{f}_n(x) = \text{argmin}_{m_x \in \mathbb{R}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(Y_i - m_x)^2$$
    I.e., at each point x, we are searching for the best local constant m_x such that the localized sum of squares is minimized
  - Solution: Nadaraya-Watson kernel estimator
    $$\hat{f}_n(x) = \sum_{i=1}^{n} w_i(x) Y_i \text{ where } w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$$

In practice, choice of kernel function does not have a strong influence on outcome, but choice of bandwidth h does. h governs *Bias-Variance tradeoff*.[31]

- bias $\sim h^2$;
- variance $\sim 1/h$

Rates of convergence:

- OLS (parametric): n^{-1}
- Non-parametric: $n^{-4/5}$

Derivation of convergence rates and h proportionality

OLS, $\hat{y} = \ell^T \beta = \hat{f}(\ell)$.

MSE equals variance since OLS is unbiased:

$$\text{Var}\left(\hat{y}\right) = \ell^T \sigma^2 (X^T X)^{-1} \ell = \ell^T \sigma^2 \left(\underbrace{\frac{X^T X}{n}}_{\to \Sigma}\right)^{-1} \ell n^{-1}$$

Here we make our usual assumption that $X^T X n^{-1}$ converges to a constant, as for instance also in Lemma 5.1.[32]

[31] MSE convergence and bias proportionality requires f to be twice continuously differentiable. Variance proportionality does not require this assumption.

[32] For each entry (j, k), we're essentially assuming that the sample covariance between predictors j and k converges to some population value.

Kernel regression

We have

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) Y_i \Big/ \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

We for now assume that $x_i = \frac{i}{n}$. i.e. the x values are on a grid with difference $\frac{1}{n}$. Then, using $\mathbb{E}(Y_i) = f(X_i)$,

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(x)\right] &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) f(i/n) \Big/ \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \\ &\approx \frac{\frac{1}{h}\int K\left(\frac{x-\mu}{h}\right) f(\mu)d\mu}{\frac{1}{h}\int K\left(\frac{x-\mu}{h}\right) d\mu}\end{aligned}$$

where we used that an integral is approximately a sum. Note that n was moved from outside the integral since the integration variable $\mu \approx \frac{1}{n}$ (loosely speaking).

We now perform a change of variables: $(x-\mu)/h = -w$. Then $d\mu = hdw$. Further, assume $K()$ is a symmetric density. Then we obtain

$$\frac{\frac{1}{h}\int K(-w) f(x+wh) h dw}{\frac{1}{h}\int K(-w) h dw}$$

Using symmetry and the fact that a density integrates to 1, simpliyfing $1/h \cdot h$ in enumerator and denominator, we get

$$= \int K(w) f(x+wh) dw$$

Using a taylor approximation, $f(x+hw) \approx f(x) + f'(x)hw + 1/2 f''(x)h^2w^2$, noting that due to twice continuous differentiability the remaining terms are small,[33] we have

$$\mathbb{E}\left[\hat{f}_h(x)\right] = f(x)\underbrace{\int K(w)dw}_{=1} + f'(x)h\underbrace{\int K(w)w dw}_{=0 \text{ (symmetry)}} + f''(x)h^2\frac{1}{2}\underbrace{\int K(w)w^2 dw}_{=:C_1(K)} + \ldots$$

Our bias is therefore, approximately,

$$\mathbb{E}\left[\hat{f}_h(x)\right] - f(x) = f''(x)h^2\frac{1}{2}\int K(w)w^2 dw$$

which is proportional to h^2. Additionally, our bias depends on curvature.[34]

Similarly, we can compute the variance. $\hat{f}_h(x)$ is a sum of independent terms since it is

[33] This follows from the rest term in the taylor expansion and continuous functions being bounded on closed intervals.

[34] If $f''(x) < 0$, e.g. at a maximum, we underestimate, if $f''(x) > 0$, e.g. at a minimum, we overestimate. This is a general feature of Kernel regression, called *errosion*.

a sum of independent Y_i, so

$$\begin{aligned}\operatorname{Var}\left(\hat{f}_h(x)\right) &= \frac{1}{n^2h^2}\sum_{i=1}^{n} K^2\left(\frac{x-x_i}{h}\right)\sigma_\varepsilon^2 \Big/ \left(\frac{1}{nh}\sum K\left(\frac{x-x_i}{h}\right)\right)^2 \\ &\approx \frac{1}{nh^2}\int K^2(w)\,h dw\,\sigma_\varepsilon^2 \Big/ \underbrace{\left(\frac{1}{h}\int K(-w)h\,dw\right)^2}_{\to 1} \\ &= \frac{1}{nh}\underbrace{\int K^2(w)dw}_{=:C_2(K)}\sigma_\varepsilon^2\end{aligned}$$

which is proportional to h^{-1}. Notice that we have one n remaining when going to the Riemann sum (integral), since we first square the $1/nh$ term when taking it out of the variance.

Therefore, we conclude:

$$MSE(\hat{f}_h(x)) = h^4(f''(x))^2C_1(K)^2 + \frac{\sigma_\varepsilon^2}{nh}C_2(K)$$

Optimizing with respect to h gives[35]

$$h_{opt} = n^{-1/5}\left(\frac{\sigma^2 C_2(K)}{4(f''(x))^2C_1(K)}\right)^{1/5}$$

Plugging back into the MSE, we get

$$MSE(\hat{f}_{h,opt}(x)) = n^{-4/5}\cdot C$$

Inference

The kernel estimator at design points is actually a linear operator. To see this, assume $Y = m(x) + \varepsilon, \mathbb{E}(\varepsilon) = 0, \operatorname{Cov}(\varepsilon) = \sigma^2 1$ and consider $(\hat{m}(x_1), \ldots, \hat{m}(x_n)) =: \hat{m}(x_0) = \hat{y}$. Since

$$\hat{m}(X) = \frac{\sum K\left(\frac{X-X_i}{h}\right)y_i}{\sum K\left(\frac{X-X_i}{h}\right)} = \sum w_i(X)y_i$$

we can write $\hat{y} = SY$ with $[S]_{rs} = w_s(x_r)$. We have

- $\mathbb{E}(\hat{m}(x_0)) = Sm(x_0) \neq m(x_0)$
- $\operatorname{Cov}(\hat{m}(x_0)) = \sigma_\varepsilon^2 SS^T$
- We may assume asymptotically $\hat{m}(x) \approx \mathcal{N}\Big(E(\hat{m}(x)), \operatorname{Var}(\hat{m}(x))\Big)$
- We estimate $\widehat{\text{s.e.}}(\hat{m}(x_i)) = \hat{\sigma}\cdot\sqrt{(SS^T)_{ii}}$ and $\hat{\sigma}^2 = \sum_{i=1}^n \frac{\left(Y_i - \hat{m}(x_i)\right)^2}{n-df}$ where $df = \operatorname{tr}(S)$.
- Then $\hat{m}(x_i) \pm 1.96\cdot\widehat{\text{s.e.}}(\hat{m}(x_i))$ is an approximate 95% CI for $\mathbb{E}(\hat{f}(x_i))$

[35] Notice that the optimal value depends on the unknown function f. Thus we cannot use this practically to choose h. Instead we use cross-validation.

Local Polynomial Regression

Instead of calculating the average within a bandwidth (with reduced weights for far away points), we may also fit a weighted local polynomial regression, centered at x:

$$\hat{\beta}(x) = \arg\min_{\beta \in R^p} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \left(Y_i - \beta_1 - \beta_2 (x_i - x) - \cdots - \beta_p (x_i - x)^{p-1}\right)^2.$$

If we use a degree 1 polynomial, this is also called a *local linear estimator* or local linear regression.

- Low order polynomials used
- Have local regression function at a point x:

$$g_x(u) = \sum_{j=1}^{p} \hat{\beta}_j(x)(u - x)^{j-1} \text{ evaluated at } u = x$$

- Use **intercept only** for prediction:[36]

$$\hat{f}_h(x) = \hat{\beta}_1(x)$$

  I.e., for any new point x', we refit the model around this point and take the newly fitted $\hat{\beta}_1(x')$ to be the prediction.

- Advantages:
  - Better performance at edges
  - Can estimate derivatives: Compute r-th derivative wrt. u and evaluate at $u = x$[37]

$$\to \hat{f}^{(r)}(x) = r!\hat{\beta}_{r+1}(x), \text{ for } r = 0, 1, \ldots, p-1$$

Smoothing Splines

Idea: Among all functions $m(x)$ with continuous second derivatives, find the one which minimizes the penalized residual sum of squares:

$$\sum_{i=1}^{n} \left(Y_i - m\left(x_i\right)\right)^2 + \lambda \int m''(z)^2 dz$$

Solution: *Natural cubic spline* with knots at sorted observations $x_1, \ldots, x_n$.[38]

- Cubic polynomials around each point x_i

[36] Note that if the design matrix is not orthogonal, this is not equal to just taking the average.

[37] Note that x is fixed and we fitted a linear function near x and evaluate it at u. Hence, to find the correct slope, we also have to take the derivative w.r.t. u and evaluate this derivative at $u = x$.

[38] Our estimated function is piecewiese cubic polynomial and thus has $4(n-1)$ parameters. Requiring continuity for function and first two derivatives imposes $3(n-2)$ inner continutiy constraints, additionally restrictions $\widehat{m}''(x_1) = \widehat{m}''(x_n) = 0$ leads to n free parameters.

- Smooth connections
- Curvature 0 at outer observations

Since the solution is a cubic spline, we can represent the solution using adequate basis functions B_j for natural splines:

$$m_\lambda(x) = \sum_{j=1}^{n} \beta_j B_j(x)$$

Optimize:

$$|Y - B\beta|^2 + \lambda \beta^T \Omega \beta,$$

where the design matrix B has columns $\left(B_j\left(x_1\right), \ldots, B_j\left(x_n\right)\right)^T$ and $\Omega_{jk} = \int B_j''(z) B_k''(z) dz$.

- Solution: $\hat{\beta} = \left(B^T B + \lambda \Omega\right)^{-1} B^T Y$
- Predictions: $\hat{Y} = B\hat{\beta} = B\left(B^T B + \lambda \Omega\right)^{-1} B^T Y = S_\lambda Y$
- Approximately similar to kernel regression with adaptive bandwidth[39]

Generalized Additive Models

In higher dimensions, close neighbors become rare and distances between all points become large. Thus, kernel regressions which depend on "neighborhoods" will fail. GAMs reduce this problem by considering non-parametric regression per dimension (at price of assuming additivity).

$$g_{add}(x) = \mu + \sum_{j=1}^{p} g_j\left(x_j\right), \mu \in \mathbb{R}$$
$$g_j(\cdot) : \mathbb{R} \to \mathbb{R},\ \mathbb{E}\left(g_j\left(X_j\right)\right) = 0 \quad (j = 1, \ldots, p)$$

7.3 High-dimensional Regression

If we have $p > n$, we have a perfect fit $\hat{y}$ and parameter estimates $\hat{\beta}$ are no longer unique. We instead impose regularisation penalties. E.g., we minimize

$$PLS(\beta) = (y - X\beta)^T (y - X\beta) + \lambda \cdot \text{pen}(\beta)$$

which we may equivalently write as

$$\min (y - X\beta)^T (y - X\beta) \quad \text{s.t.} \quad \text{pen}(\beta) \leq s$$

Even if $p \leq n$, such a setup may be useful due to the bias-variance tradeoff as a biased model may outperform OLS w.r.t. prediction accuracy (see (3.4)).

[39] In regions where data suggests more curvature (higher second derivatives), the smoothing spline effectively decreases its bandwidth to capture these features

Ridge Regression

$$(y - X\beta)^T(y - X\beta) + \lambda \cdot \sum_{j=1}^{p} \beta_j^2$$

- Do not penalize intercept
- "Shrinkage penalty" λ is a tuning parameter, find with cross-validation
- Standardise predictors
- Closed form solution:
$$\hat{\beta}_{\text{Ridge}} = (X^T X + \lambda \mathbf{1})^{-1} X^T y$$
- Note that $X^T X + \lambda \mathbf{1}$ is always pd whenever $\lambda > 0$.
  - Thus, $\hat{\beta}_{\text{Ridge}}$ always exists and is always unique.
  - All p coefficients non-zero even if $p \gg n$
- Covariance matrix: $\text{Cov}\left(\hat{\beta}_{\text{idge}}\right) = \sigma^2 \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{X} \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1}$
- $\text{Cov}\left(\hat{\beta}_{\text{LS}}\right) - \text{Cov}\left(\hat{\beta}_{\text{Ridge}}\right)$ is p.d. for $\lambda > 0$ so Ridge coefficients have a smaller variance
- As $\lambda \to \infty$, the L_2 norm of $\hat{\beta}$ decreases monotonically, but individual coefficients might have a bump and increase.[40]
- If $p > n$, as $\lambda \to 0$, Ridge bias does not go to zero. If $n \leq p$, bias goes to zero since $\lim_{\lambda \to 0} \hat{\beta}_{\text{Ridge}}(\lambda) = \left(X^T X\right)^{-1} X^T Y = \hat{\beta}_{\text{OLS}}$.

Using penalisation introduces bias but may lead to better out-of-sample prediction performance due to bias-variance tradeoff.

LASSO

Issue with ridge: still have all variables in model. Advantage of LASSO: Shrink some coefficients to zero, thus perform variable selection.[41]

Objective function:

$$\min (y - X\beta)^T(y - X\beta) + \lambda \cdot \sum_{j=1}^{p} |\beta_j|$$

[40] If the design matrix is orthogonal, the decrease is continous and

$$\hat{\beta}_{\text{Ridge},j}(\lambda) = \frac{1}{1+\lambda} \hat{\beta}_{\text{LS},j}.$$

[41] We can view these as approximations of the NP-hard problem:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I\left(\beta_j \neq 0\right) \leq s.$$

- ℓ_1 norm not differentiable, no closed-form solution, non-linear estimator
- Still convex optimization problem
- At most $\min\{n, p\}$ non-zero coefficients
- Many theoretical advantages if true model is sparse
- In case of orthonormal design matrix, one can show that

$$\hat{\beta}_{\text{LASSO},j}(\lambda) = \text{sign}\left(\hat{\beta}_{\text{LS},j}\right)\left[\left|\hat{\beta}_{\text{LS},j}\right| - \frac{\lambda}{2}\right]_+$$

Theoretical Advantages of LASSO

Under technical assumptions (true model sparsity, relation between predictors, ...):

- If we knew the true non-zero coefficients β^0, MSE of OLS would converge as:

$$\frac{\left|X\left(\widehat{\beta} - \beta^0\right)\right|^2}{n} = O_P\left(\frac{s_0}{n}\right)$$

  where s_0 is number of non-zero coefficients in true model.

- MSE convergence of LASSO:

$$\frac{\left|X\left(\widehat{\beta} - \beta^0\right)\right|^2}{n} = O_P\left(\frac{s_0 \log(p)}{n}\right)$$

  We thus pay a small price of using LASSO and not knowing true non-zero coefficients

- LASSO consistently estimates parameters:

$$\left|\hat{\beta} - \beta^0\right|_2 = O_P\left(\sqrt{\frac{s_0 \log(p)}{n}}\right)$$

  where β^0 is the vector of active βs (padded with 0s).

Issues in High-Dimensional Statistics

If we have $p \geq n$, OLS overfits. Fit on training data will be exact while out-of-sample testing error gets worse the more variables are added. Cannot use BIC, AIC, etc. as we cannot estimate $\hat{\sigma}^2$.

Penalisation helps tame the hamrful flexibiltiy leading to better predictions in high dimensions.

- Inference:

  - Cannot use traditional measures such as p-values, training MSE due to perfect fit in-sample
  - Instead focus on out-of-sample predictive performance
  - Advanced in-sample methods exist
- Model interpretation:
  - Perfect multicolinearity is guaranteed if $p > n$.
  - Any model could be replaced by other model with correlated predictors
  - Interpret model as one of many possible (good fitting) models

A Mathematical Statistics

A.1 Fisher Information and Fisher Scoring

(From DeGroot & Shervish)

Definition A.1 Fisher Information

Let X be a random variable whose distribution depends on a parameter θ that takes values in an open interval Ω of the real line. Let the p.f. or p.d.f. of X be $f(x \mid \theta)$. Assume that the set of x such that $f(x \mid \theta) > 0$ is the same for all θ and that $\lambda(x \mid \theta) = \log f(x \mid \theta)$ is twice differentiable as a function of θ. The **Fisher information** $I(\theta)$ in the random variable X is defined as

$$I(\theta) = E_\theta \left\{ \left[\lambda'(X \mid \theta) \right]^2 \right\}.$$

Thus, if $f(x \mid \theta)$ is a p.d.f., then

$$I(\theta) = \int_S \left[\lambda'(x \mid \theta) \right]^2 f(x \mid \theta) dx.$$

The idea behind Fischer information is that if the density f is sharply peaked with respect to θ, then it is easier to find the 'correct' value of θ. We therefore use the *score* (i.e., the derivative of the log of the density with respect to the parameter). A random variable carrying high Fisher information implies that the absolute value of the score is often high. High Fischer information indicates that the maximum is 'sharp' while low Fischer information means that the maximum is 'blunt'.

When there are N parameters, so that θ is an $N \times 1$ vector $\theta = \begin{bmatrix} \theta_1 & \theta_2 & \dots & \theta_N \end{bmatrix}^\top$, the Fisher information takes the form of an $N \times N$ matrix. The typical element of this matrix is given by:

$$[\mathcal{I}(\theta)]_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X; \theta) \right) \left(\frac{\partial}{\partial \theta_j} \log f(X; \theta) \right) \middle| \, \theta \right].$$

Under certain regularity conditions, the Fisher information matrix may also be written as

$$[\mathcal{I}(\theta)]_{i,j} = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(X; \theta) \middle| \, \theta \right]. \tag{A.1}$$

We may use the Fischer information matrix to perform parameter estimation. This is known as **Fisher scoring.** Suppose we are given a log-likelihood function $\ell(\theta)$ and we want to maximize with respect to the parameter θ. We may use the scores (i.e., derivative of log-likelihood functions) and the Hessian matrix $\mathcal{J}$ to optimize this using Newton-Raphson:

$$\theta_{t+1} = \theta_t - \left[\mathcal{J}(\theta_t) \right]^{-1} \nabla \ell(\theta_t)$$

If we instead use the expectation of the negative Hessian, which is, under certain regularity conditions the Fisher information matrix given in (A.1), we do Fischer scoring.

$$\theta_{t+1} = \theta_t + \left[\mathcal{I}(\theta_t) \right]^{-1} \nabla \ell(\theta_t)$$

A.2 Convergence

We introduce different notions of convergene of random variables.

Definition A.2 Convergence in Distribution

A sequence $X_1, X_2, \ldots$ of real-valued random variables, with cumulative distribution functions $F_1, F_2, \ldots$, is said to **converge in distribution**, or *converge weakly*, or converge in law to a random variable X with cumulative distribution function F if

$$\lim_{n \to \infty} F_n(x) = F(x),$$

for every number $x \in \mathbb{R}$ at which F is continuous.

- Convgerence in distibution implies that the probability for X_n to be in a given range is approximately equal to the probability of X being in that range for n sufficiently large
- Convergence in distribubtion does not generally imply that the sequence of densities will convergence
- There exist many equivalent definitions

Definition A.3 Convergence in Probabilty

A sequence $\{X_n\}$ of random variables **converges in probability** towards the random variable X if for all $\varepsilon > 0$

$$\lim_{n \to \infty} \mathbb{P}\left(|X_n - X| > \varepsilon\right) = 0$$

More explicitly, let $P_n(\varepsilon)$ be the probability that X_n is outside the ball of radius ε centered at X. Then X_n is said to converge in probability to X if for any $\varepsilon > 0$ and any $\delta > 0$ there exists a number N (which may depend on ε and δ) such that for all $n \geq N, P_n(\varepsilon) < \delta$,

- Idea: probability of *unusual* outcomes becomes smaller as n increases.
- Convergence in probability implies convergence in distribution

Convergence in probability is closely related to an important characteristic of many estimators:

Definition A.4 Consistent estimator

An estimator T_n of parameter θ is said to be **weakly consistent**, if it converges in probability to the true value of the parameter:

$$\text{plim}_{n \to \infty} T_n = \theta,$$

i.e. if, for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\left(|T_n - \theta| > \varepsilon\right) = 0.$$

A very powerful theorem is the following which tells us that usual operations such as additiona of random variables has a nice "consistency" property. Part (3) of the theorem allows us to replace covariances with estimated covariances for asymptotic distributions of test statistics.

Theorem A.1 Slutsky's theorem

Let X_n, Y_n be sequences of scalar/vector/matrix random elements. If X_n converges in distribution to a random element X and Y_n converges in probability to a constant c, then

1. $X_n + Y_n \xrightarrow{d} X + c$;
2. $X_n Y_n \xrightarrow{d} Xc$;
3. $X_n / Y_n \xrightarrow{d} X/c$, provided that c is invertible,

where $\xrightarrow{d}$ denotes convergence in distribution.