Chapter 1: Multiple Linear Regression

The general linear model

\[ Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j} + \epsilon_i \]

With:

- \( Y_i \) the \( i \)th observation of the **response** variable
  - dependent variable
  - explained, observed variable
- \( x_{i,j} \) the (observed) \( i \)th value of the \( j \)th **covariate**
  - independent variable
  - regression variable
  - explanatory variable
  - predictor, predicting variable
  - control variable
- \( \beta_j \) the regression coefficient corresponding to the \( j \)th covariate
  - this is a **parameter** vector to be estimated
- \( \beta_0 \) the intercept (other coefficients sometimes “slopes”)
- \( \epsilon_i \) the error (observational noise) in the \( i \)th observation

In matrix form:

\[ Y = X\beta + \epsilon \]

where \( X \in \mathbb{R}^{n \times p} \) and the first column of \( X \) all ones, \((X_{i,0} = 1)\) corresponding to the intercept. The other columns of \( X \) are observed covariate values.

---

Examples, specific cases

1. **Polynomial regression**

\[ Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{i,j}^j + \epsilon_i \]

So \( x_{i,j} = x_{i,j}^j \).

There is only one explanatory variable in strict sense (\( x \) with observations \( x_i \)), but the response depends on several powers of \( x \).

2. **Models with interaction**

Suppose: 2 explanatory variables, \( x_1 \) and \( x_2 \), but \( Y \) also depends on the interaction between them

\[ Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_{1,2} x_{i,1} x_{i,2} \]

We then set \( x_3 = x_1 x_2 \).

3. **One-way variance analysis (ANOVA)**

\[ Y_{ij} = \mu_i + \epsilon_{ij} \]

with \( i = 1, \ldots, k \) and \( j = 1, \ldots, n_i \).
Note that in this model the observations have double indices. All observations at the same level \( \mu_i \) are only different by the observational noise, so in a regression model they should share common variables of the explanatory variables.

The model becomes

\[
Y_{ij} = \sum_{\ell=1}^{k} x_{ij,\ell} \mu_{\ell} + \varepsilon_{ij}
\]

where we take indices \( i \) and \( j \) together into a single index \( ij \) and \( x_{ij,\ell} \) gets the value 1 if \( \ell = i \) and 0 otherwise.

### The normal equation

Model \( Y = X\beta + \varepsilon \)

We search for \( \hat{\beta} \) so that \( X\hat{\beta} \) is as close as possible to \( Y \) in quadratic norm.

Algebraically this means that we look for the minimum least squares solution to the overdetermined set of equations \( Y = X\beta \)

\[
\| Y - X\beta \| \text{ has to be minimized. In a Euclidian space (i.e., a space where the notion of orthogonality can be defined using inner products) that means that the residual } Y - X\hat{\beta} \text{ has to be orthogonal (perpendicular) to the approximation (or projection) } X\hat{\beta} \text{ in the subspace spanned by } X.
\]

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

This is the case if \( X^T(Y - X\hat{\beta}) = 0 \)

---

### Orthogonalisation

If \( X \) has orthogonal columns, then \( X^T X = I_p \), and thus the projection becomes

\[
\hat{y} = X \cdot X^T \cdot Y
\]

The \( i \)th component is then

\[
\hat{y}_i = \sum_{j=0}^{n-1} x_{i,j} x_{k,j} \cdot Y_k
\]

The computation of all \( n \) components of \( \hat{y} \) can proceed in parallel, without numerical errors progressing from one component to another.

Otherwise, we Orthogonalise \( X \)
Gram-Schmidt-orthogonalisation

Let \( x_j, j = 0, \ldots, p-1 \) be the columns of \( X \), then define

\[
q_0 = x_0 = 1 \\
q_1 = x_1 - \frac{\langle x_1, q_0 \rangle}{\|q_0\|^2} \cdot q_0 \\
\vdots \\
q_j = x_j - \sum_{k=0}^{j-1} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

This is:

- \( q_j \) is the residual of orthogonal projection of \( x_j \) onto \( q_0, \ldots, q_{j-1} \)
- Looks like linear regression within the covariate matrix \( X \)
- The new covariates \( q_j \) are orthogonal, but not orthonormal (they can be easily normalised)

Regression after orthogonalisation

Plugging into the solution of the normal equation:

\[
\hat{y} = QR(R^TQ^TQ)^{-1}R^TQ^T \cdot Y = Q(Q^TQ)^{-1}Q^T \cdot Y
\]

This is a projection onto the orthogonal (but not orthonormal) basis \( Q \).

Elaborated

\[
\hat{y} = \sum_{k=0}^{p-1} \frac{\langle Y, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

The vector of coefficients \( \hat{\gamma} \) with

\[
\hat{\gamma}_k = \frac{\langle Y, q_k \rangle}{\|q_k\|^2}
\]

satisfies \( \hat{y} = Q\hat{\gamma} \) and \( \hat{\gamma} = (Q^TQ)^{-1}Q^T \cdot Y \)

Gram-Schmidt in matrix form

We have

\[
x_j = q_j + \sum_{k=0}^{j-1} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

Apply scalar product with \( q_j \) on both sides to find that \( \frac{\langle x_j, q_j \rangle}{\|q_j\|^2} = 1 \),

Hence

\[
x_j = \sum_{k=0}^{j} \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \cdot q_k
\]

In matrix-form:

- **QR-decomposition** \( X = QR \) with the columns \( Q \) equal to \( q_j \), and \( R \) an upper triangular matrix, with entries \( R_{k,j} = \frac{\langle x_j, q_k \rangle}{\|q_k\|^2} \).

\( R \in \mathbb{R}^{p \times p} \) and \( R \) is invertible.

**Interpretation** Every column of \( X \) (every covariate \( x_j \) can be written as a linear combination of orthogonal covariates, which are the columns of \( Q \).

(Right-multiplication with \( R \))

Relation between \( \hat{\gamma} \) and \( \hat{\beta} \):

\[
\hat{\beta} = (X^TX)^{-1}X^T \cdot Y = (R^TQ^TQ)^{-1}R^TQ^T \cdot Y = R^{-1}(Q^TQ)^{-1}Q^T \cdot Y
\]

\[
\hat{\beta} = R^{-1}\hat{\gamma} \quad \hat{\gamma} = R\hat{\beta} \quad \text{(Left-multiplication with } R \text{)}
\]
Elaboration for $p = 1$, simple linear regression

Model

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n \]

Simple linear regression: $x_0 = 1$ and $x_1 = x$.

\[ q_1 = x_1 - \frac{\langle x_1, q_0 \rangle}{\|q_0\|^2} \cdot q_0 \]
\[ = x_1 - \frac{\langle x_1, 1 \rangle}{\|1\|^2} \cdot 1 \]
\[ = x_1 - \frac{\sum_{i=1}^n x_{i,1}}{n} \cdot 1 = x_1 - \overline{x}_1 \cdot 1 \]

\[
R = \begin{bmatrix}
\frac{1}{n} & \overline{x}_1 \\
0 & 1
\end{bmatrix}
R^{-1} = \begin{bmatrix}
\frac{n}{n} & -n \overline{x}_1 \\
0 & 1
\end{bmatrix}
\]

\[ \hat{\gamma}_0 = \overline{Y} \]
\[ \hat{\gamma}_1 = \frac{\langle q_1, Y \rangle}{\|q_1\|^2} = \frac{\langle x_1 - \overline{x}_1, 1, Y \rangle}{\|x_1 - \overline{x}_1, 1\|^2} \]

Since the second row of $R^{-1}$ equals $[0 \ 1]$, we find immediately that

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(x_{i,1} - \overline{x}_1)}{\sum_{i=1}^n (x_{i,1} - \overline{x}_1)^2} \]
\[ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}_1 \]

Tools for statistical exploration

In order to further explore the properties of the least squares estimator, we need to establish some classical results in multivariate statistics

1. Covariance matrix
2. Multivariate normal distribution
3. Trace of a matrix

Covariance matrix

Let $X$ be a $p$-variate random vector with joint density function $f_X(x)$, joint (cumulative) distribution function $F_X(x)$, and joint characteristic function $\phi_X(t) = E(e^{it^TX})$.

\[ \mu = E(X) \]
\[ \Sigma_X = E[(X - \mu)(X - \mu)^T] \]

We have $(\Sigma_X)_{ij} = cov(X_i, X_j)$

\[ \Sigma_X \] is symmetric

If $Y = AX$ Then $\Sigma_Y = A \Sigma_X A^T$

Proof:

It holds that $\mu_Y = A \mu_X$, so

\[ \Sigma_Y = E[(Y - \mu_Y)(Y - \mu_Y)^T] = E[(A(X - \mu_X))(A(X - \mu_X))^T] \]

So, with $A_i$ the $i$th column of $A$, we have
\[ \text{cov}(Y_i, Y_j) = A_j^T \Sigma X A_j = A_i^T \Sigma X A_i \]

A symmetric matrix \( S \) is called positive (semi-)definite if \( x^T S x \geq 0 \) for every vector \( x \neq 0 \).

A positive-definite matrix always has real, positive eigenvalues.

The multivariate normal distribution

**Definition: multivariate normal distribution**

\[ \mathbf{X} \sim N(\mu, \Sigma_X) \iff Z_i \sim N(0, 1) \]

As the components of \( Z \) are uncorrelated, and (here) normally distributed, they must be mutually independent.

**The joint density function**

\[ f_X(x) = \frac{1}{(2\pi)^{p/2} \sqrt{\det(\Sigma_X)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma_X^{-1}(x-\mu)} \]

**The characteristic function**

\[ \phi_X(t) = \exp \left( it^T \mu - \frac{1}{2} t^T \Sigma_X t \right) \]

For this matrix it holds \( |\det(A)|^2 = \det(\Sigma_X) \)

Suppose now \( Z = A^{-1} (X - \mu) \) (with \( \mu = EX \)), then \( EZ = 0 \) and \( \Sigma_Z = A^{-1} (AA^T) A^{-T} = I \)

The components of \( Z \) are thus uncorrelated.

**Vice versa**, if \( \Sigma_Z = I \) and \( A \) is an arbitrary matrix, and \( X = AZ \), then \( \Sigma_X = AA^T \) and this arbitrary \( A \) can then be written as \( A = T \Lambda^{1/2} \), where \( T \) orthogonal.

**Properties of the multivariate normal distribution**

- **Property 1** \( X \sim (\mu, \Sigma_X) \), then \( (X - \mu)^T \Sigma_X^{-1} (X - \mu) \sim \chi^2 \)

This follows from

\[ (X-\mu)^T \Sigma_X^{-1} (X-\mu) = (X-\mu)^T A^{-T} A^{-1} (X-\mu) = Z^T Z = \sum_{i=1}^p Z_i^2 \]

and \( Z_i^2 \sim \chi^2_1 \) and \( Z_i \) mutually independent

- **Property 2**

\[ X \sim N(\mu, \Sigma_X) \]

If \( X \sim N(\mu, \Sigma_X) \), then for \( C \in \mathbb{R}^{p \times q} \) with \( q \leq p \) \( CX \sim N(C\mu, C\Sigma_X C^T) \)

- **Property 3** \( X \sim (\mu, \Sigma_X) \leftrightarrow a^T X \sim N(a^T \mu, a^T \Sigma_X a) \)
Properties of the multivariate normal distribution

**Property 4**

Suppose \( X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \) and \( X \sim N(\mu, \Sigma_X) \) with \( \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \) and

\[
\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\]

and \( \det(\Sigma_X) \neq 0 \), then

\[
X_1 | X_2 = x_2 \sim N(\mu_{1|2}, \Sigma_{1|2})
\]

with \( \mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \) and \( \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \)

It holds that

\[
\Sigma_{-1}^{1} = (\Sigma_{-1}^{1})_{11}
\]

Note also that \( \Sigma_{21} = \Sigma_{12} \).

Marginal and conditional covariance

(Elaboration of previous slide)

Suppose \( X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \) with mean \( \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \) and covariance matrix

\[
\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}
\]

and \( \det(\Sigma_{22}) \neq 0 \), then

**Marginal covariance**

\[
\text{cov}(X_1) = \Sigma_{11}
\]

**Conditional covariance**

\[
\text{cov}(X_1 | X_2) = \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}
\]

This expression is known as the Schur-complement.

It holds that

\[
\text{cov}(X_1 | X_2) = \Sigma_{1|2} = [ (\Sigma_{-1}^{1})_{11} ]^{-1}
\]

Trace of a matrix

\[
\text{Tr}(A) = \sum_{i=1}^{p} A_{ii}
\]

\[
\text{Tr}(AB) = \text{Tr}(BA) \quad (A \in \mathbb{R}^{p \times q}, B \in \mathbb{R}^{q \times p})
\]

Holds for both square \((p = q)\) and rectangular matrices with matching sizes

Proof:

\[
\text{Tr}(AB) = \sum_{i=1}^{p} (AB)_{ii} = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{ij} B_{ji} = \sum_{j=1}^{q} \sum_{i=1}^{p} B_{ji} A_{ij} = \sum_{j=1}^{q} (BA)_{jj} = \text{Tr}(BA)
\]

**Corollary** A, B, C square matrices, then:

\[
\text{Tr}(ABC) = \text{Tr}(BCA) = \text{Tr}(CAB)
\]

Trace remains unchanged under cyclic permutation, so NOT under every permutation:

\[
\text{Tr}(ABC) \neq \text{Tr}(BAC)
\]

Holds also for more than 3 matrices and also for rectangular matrices with matching dimensions

\[
\text{Tr}(A + B) = \text{Tr}(A) + \text{Tr}(B) \quad \text{Attention: Tr}(A \cdot B) \neq \text{Tr}(A) \cdot \text{Tr}(B)
\]
Maximum likelihood for regression

**Model** \( Y = X \beta + \varepsilon \) (random variable \( Y \) instead of \( X \) in previous slides)

**Assume** \( \varepsilon \sim \text{N.I.D.}(0, \sigma^2) \), then joint observational density is multivariate normal, more precisely

\[
L(\beta, \sigma^2; Y) = \frac{1}{(2\pi \sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)}
\]

To be maximised over \( \beta \) and \( \sigma^2 \).

\[
\log L(\beta, \sigma^2; Y) = -\frac{n}{2} \log (2\pi \sigma^2) - \frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta)
\]

Dependence on \( \beta \) only through second term, so \( \beta \) maximises \( L(\beta, \sigma^2; Y) \) or \( \log L(\beta, \sigma^2; Y) \) iff \( \beta \) minimises

\[
(Y - X\beta)^T (Y - X\beta) = ||Y - X\beta||^2
\]

This is least squares.

Conclusion: for normal errors is the least squares solution the maximum likelihood solution.

---

Properties of maximum likelihood estimators

**Model** \( Y = X \beta + \varepsilon \)

**Least squares:** \( \hat{\beta} = (X^T X)^{-1} X^T Y \)

**Property 1: unbiased**

\[
E\hat{\beta} = \beta
\]

**Property 2: covariance**

\[
\Sigma_\beta = \sigma^2 (X^T X)^{-1}
\]

**Property 3: Normality** \( \beta \) is normally distributed (indeed, the estimator is a linear combination of normally distributed random variables)

**Marginal distributions** \( \hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}) \)

---

Property 4: best linear unbiased estimator (BLUE)

Denote \( y = X \beta \) and \( \hat{y} = X \hat{\beta} = X (X^T X)^{-1} X^T Y = PY \)

We consider estimators for linear combinations of the expected responses: \( \hat{\theta} = c^T \hat{y} = c^T X \beta \).

For instance the \( \beta_j \)'s themselves. Indeed: \( \beta = (X^T X)^{-1} X^T y \). So take for \( c \) the \( j \)th column of \( X (X^T X)^{-1} \) \( c^T \) is then the \( j \)th row of \( (X^T X)^{-1} X^T \)

The LS estimator is a linear combination of the observations \( \hat{\theta} = c^T P Y = c_P^T Y \) with \( c_P = P^T c \)

**Note** \( P^T = P \) and \( PP = P \) and \( Py = y \)

We have: \( E\hat{\theta} = \theta \), so \( c^T P y = \theta = c^T y \) \( \rightarrow \) Or: \( c_P^T y = c^T y \)

Suppose now that \( \hat{\theta} = d^T Y \) were also unbiased, i.e., \( d^T y = c^T y \)

Example: denote \( X_p \) the \( p \times p \) upper part of the \( n \times p \) matrix \( X \) (that

is: restrict to the first \( p \) observations) and \( d^T = c^T X_p^{-1} \)

Because \( y = X \beta \), this becomes: \( d^T X \beta = c^T X \beta \)

\( X \) does not depend on \( \beta \), the same argument can be repeated for the same \( d \) and \( c \) and always other \( \beta \), so it follows: \( (d^T - c^T) X = 0 \)

Taking transposes: \( X^T (d - c) = 0 \)

Pre-multiplication by \( (X^T X)^{-1} \): \( P (d - c) = 0 \)

So \( P d = P c \)

And further \( \text{var}(\hat{\theta}) - \text{var}(\hat{\theta}) = \text{var}(d^T Y) - \text{var}(c^T P Y) = \text{var}(d^T Y) - \text{var}(d^T PY) = d^T d_\sigma^2 - d^T P d_\sigma^2 = d^T (I - P) d_\sigma^2 = \left\| (I - P) d \right\| \sigma^2 \geq 0 \)
Variance estimation

\[
S^2 = \frac{(Y - X\hat{\beta})^T(Y - X\hat{\beta})}{n - p}
\]

Note that the first factor is transpose (↔ expression of covariance matrix)

**Proof**

We know \( \hat{y} = X\hat{\beta} = PY \), so

\[
Y - X\hat{\beta} = (I - P)Y = (I - P)(X\beta + \varepsilon) = (I - P)\varepsilon
\]

and so

\[
E \left( (Y - X\hat{\beta})^T(Y - X\hat{\beta}) \right) = E \left[ \varepsilon^T(I - P)^T(I - P)\varepsilon \right] = E \left[ \sum_{i=1}^n \sum_{j=1}^n \varepsilon_i(I - P)_{ij}\varepsilon_j \right] = \sum_{i=1}^n (I - P)_{ii}E\varepsilon_i^2 = \text{Tr}(I - P)\sigma^2 = (n - p)\sigma^2
\]

**Theorem**

\( (n - p)S^2/\sigma^2 \sim \chi^2_{n-p} \)

- The matrix \( I - P \) is idempotent: \( (I - P)(I - P) = I - P \)
- This matrix has rank \( n - p \): all linear combinations of the columns of \( X \) belong to the kernel of the matrix, since \( (I - P)X = 0 \)
- An idempotent matrix with rank \( k \) has \( k \) independent eigen vectors with eigen value 1 and \( n - k \) independent eigen vectors with eigen value 0. The latter are the columns of \( X \)
- The matrix \( I - P \) is symmetric, so eigen vectors corresponding to eigen value 1 are orthogonal to eigen vectors with eigen value 0. Within each of the two groups of eigen vectors, orthogalisation (using Gram-Schmidt) is straightforward.

**Hypothesis testing - likelihood ratio**

We want to test if \( H_0: A\beta = c \) with \( A \in \mathbb{R}^{q \times p} \) where \( q \leq p \) and rank of \( A \) equal to its number of rows \( q \)

**Example** \( c = 0 \) and the rows of \( A = \begin{bmatrix} I & 0 \end{bmatrix} \). This leads to the hypothesis test if the first \( q \) \( \beta_j \) equal zero (other subset of \( \beta_j \) are possible)

As \( \sigma \) is unknown, it is a component in the parameter vector \( \theta = (\beta, \sigma^2) \) to be estimated.

Let \( \Omega = \{ \theta \} \) be the vector space of all values that the parameter vector can take and \( \Omega_0 = \{ \theta \in \Omega | A\beta = c \} \) the subspace of values that satisfy the null hypothesis.

**Likelihood ratio**

\[
\ell = \frac{\max_{\theta \in \Omega_0} L(\theta; Y)}{\max_{\theta \in \Omega} L(\theta; Y)}
\]
**Likelihood ratio**

**Theorem** (from mathematical statistics)

If \( n \to \infty \), then
\[
-2 \log \ell \xrightarrow{d} \chi^2_{\nu=\nu_0}
\]
\( \nu \) is the dimension of the space \( \Omega \), \( \nu_0 \) is the dimension of \( \Omega_0 \): in our case \( \nu = p, \nu_0 = p - q \)

**Maximum likelihood estimator in multiple regression**

\[
L(\beta, \sigma^2; Y) = \frac{1}{(2\pi \sigma^2)^{n/2}} e^{-\frac{1}{2} \left( Y - X \beta \right)^T (Y - X \beta)}
\]
Taking derivatives for \( \sigma^2 \) yields
\[
\hat{\sigma}^2 = \frac{1}{n} (Y - X \hat{\beta})^T (Y - X \hat{\beta}) = \frac{n p}{n} S^2
\]
where \( \hat{\beta} \) follows from taking derivatives for \( \beta_j \) (at least squares as discussed before; but we don’t need this for the forthcoming analysis)
Plugging in leads to
\[
L(\hat{\beta}, \hat{\sigma}^2; Y) = \frac{1}{(2\pi \hat{\sigma}^2)^{n/2}} e^{-n/2}
\]

---

**Estimators under reduced model**

Some heavy calculations ...
Starting from the reduced model \( H_0 : A \beta = c \) with \( A \in \mathbb{R}^{q \times p} \) where \( q < p \) and rank of \( A \) equal to number of rows \( q \)
If \( H_0 \) holds, then
\[
\begin{align*}
\hat{\beta}_0 &= \hat{\beta} + (X^T X)^{-1} A^T [A(X^T X)^{-1} A^T]^{-1} (c - A \hat{\beta}) \\
\hat{\sigma}_0^2 &= \frac{1}{n} (Y - X \hat{\beta}_0)^T (Y - X \hat{\beta}_0)
\end{align*}
\]
Denote \( \bar{y} = X \hat{\beta} \) and \( \bar{\beta}_0 = X \hat{\beta}_0 \), then
\[
\begin{align*}
\bar{y}_0 - \bar{y} &= X(\bar{\beta}_0 - \hat{\beta}) = X(\bar{\beta}_0 - \hat{\beta}) \left[ A(X^T X)^{-1} A^T \right]^{-1} (c - A \hat{\beta})
\end{align*}
\]
and
\[
(Y - \bar{y})^T (\bar{y}_0 - \bar{y}) = \left[ (I - X(X^T X)^{-1} X^T) Y \right]^T (\bar{y}_0 - \bar{y})
\]

---

**Maximum likelihood estimator within \( \Omega_0 \)**

Maximise \( L(\beta, \sigma^2; Y) \) under the restriction that \( A \beta = c \)

Technique: Lagrange multipliers (elaboration beyond the scope of this text)
leads to the same expression as before for \( \hat{\sigma}_0^2 \):
\[
\hat{\sigma}_0^2 = \frac{1}{n} (Y - X \hat{\beta}_0)^T (Y - X \hat{\beta}_0)
\]
And so:
\[
L(\hat{\beta}_0, \hat{\sigma}_0^2; Y) = \frac{1}{(2\pi \hat{\sigma}_0^2)^{n/2}} e^{-n/2}
\]
but \( \hat{\beta}_0 \) is now:
\[
\hat{\beta}_0 = \hat{\beta} + (X^T X)^{-1} A^T \left[ A(X^T X)^{-1} A^T \right]^{-1} (c - A \hat{\beta})
\]
The likelihood ratio is then
\[
\ell = \left( \frac{\hat{\sigma}_0^2}{\sigma^2} \right)^{n/2}
\]
Large values of \(-2 \log \ell\), i.e., small values of \( \ell \) lead to rejection of null hypothesis.

---

For \( Y - \bar{y}_0 = (Y - \bar{y}) + (\bar{y}_0 - \bar{y})^2 \) we have
\[
||Y - \bar{y}_0||^2 = ||Y - \bar{y}||^2 + ||\bar{y}_0 - \bar{y}||^2
\]

So
\[
\hat{\sigma}_0^2 = \frac{1}{n} ||Y - \bar{y}_0||^2
\]
\[
= \frac{1}{n} ||Y - \bar{y}||^2 + \frac{1}{n} ||\bar{y}_0 - \bar{y}||^2
\]
\[ \hat{\sigma}^2 + \frac{1}{n} \|X(X^TX)^{-1}A^T [A(X^TX)^{-1}A^T]^{-1} (c - A\hat{\beta})\|^2 = \hat{\sigma}^2 + \frac{1}{n} (c - A\hat{\beta})^T [A(X^TX)^{-1}A^T]^{-1} A(X^TX)^{-1} X^T X (X^TX)^{-1} A^T [A(X^TX)^{-1} A^T (c - A\hat{\beta})] \]

\[ = \hat{\sigma}^2 + \frac{1}{n} (c - A\hat{\beta})^T [A(X^TX)^{-1}A^T]^{-1} [A(X^TX)^{-1}A^T]^{-1} (c - A\hat{\beta}) \]

So
\[ \sigma_0^2 - \hat{\sigma}^2 = \frac{1}{n} (c - A\hat{\beta})^T [A(X^TX)^{-1}A^T]^{-1} (c - A\hat{\beta}) \]

### The proof for the F-test

We have:

1. \[ \hat{\sigma}^2 = \frac{n-p}{n} S^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n} \]

2. \[ \hat{\sigma}_0^2 - \hat{\sigma}^2 = (A\hat{\beta} - c)^T [A(X^TX)^{-1}A^T]^{-1} (A\hat{\beta} - c)/n \]

We have to prove that:

1. \[ n\hat{\sigma}^2/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi_{n-p}^2 \] (done before)

2. under \( H_0 \):

   \[ n(\hat{\sigma}_0^2 - \hat{\sigma}^2)/\sigma^2 = (A\hat{\beta} - c)^T [A(X^TX)^{-1}A^T]^{-1} (A\hat{\beta} - c)/\sigma^2 \sim \chi_q^2 \]

3. \( S^2 \) and \( \beta \) are independent

Therefore, all functions of \( S^2 \) are also independent from all functions of \( \hat{\beta} \). More precisely \( \hat{\sigma}^2 \) and \( \hat{\sigma}_0^2 - \hat{\sigma}^2 \) are mutually independent.

---

### An equivalent F-test

The distribution of \( -2 \log \ell \) is only asymptotically \( \chi^2 \). The elaboration of \( -2 \log \ell \) leads to logarithms of sample variances. Intuitively, a ratio of sample variances suggests the use of an F-test, which is exact, and without the need of taking a log.

\[ \sim \text{ANOVA} \]

Is the “extra sum of squares” (the part of the variance that is explained by giving up the restrictions imposed by the null hypothesis) significant?

Formally: \[ F = \frac{(\hat{\sigma}_0^2 - \hat{\sigma}^2)/q}{\hat{\sigma}^2/(n-p)} \]

We have: \[ F = \frac{n-p}{q} \frac{\hat{\sigma}_0^2 - \hat{\sigma}^2}{\hat{\sigma}^2} = \frac{n-p}{q} (\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2} - 1) = \frac{n-p}{q} (\ell - n/2 - 1) \]

We now show that \( F \sim F_{q, n-p} \)
The proof for the $F$-test: independence $S^2$ and $\hat{\beta}$

We show that $(Y - X\hat{\beta})$ is independent from $\hat{\beta}$. As both are normally distributed, it suffices to show that they are uncorrelated.

The covariance matrix  

\[
\text{cov}(Y - X\hat{\beta}, \hat{\beta}) = E \left[ (Y - X\hat{\beta})(\hat{\beta})^T \right] - E(Y - X\hat{\beta}) \cdot E(\hat{\beta})^T \\
= E \left[ (Y - X(X^TX)^{-1}X^T Y)(X^TX)^{-1}X^TY)^T \right] - 0 \cdot \beta^T \\
= (I - X(X^TX)^{-1}X^T) \cdot E(YY^T) \cdot (X^TX)^{-1}X^TY^T \\
= (I - X(X^TX)^{-1}X^T) \cdot (\sigma^2 I + X \beta \beta^T X^T) X (X^TX)^{-1} \\
= (I - X(X^TX)^{-1}X^T) \cdot (X^TX)^{-1} \sigma^2 + X \beta \beta^T \\
= 0
\]

So $P \left( \frac{|c^T(\hat{\beta} - \beta)|}{S \sqrt{c^T(X^TX)^{-1}c}} \geq t_{n-p,\alpha/2} \right) = \alpha$

We verify now that the simultaneous confidence interval follows indeed from the $F$ test for the full model.

We search for $M_\alpha$ so that $P \left( \max_c \frac{|c^T(\hat{\beta} - \beta)|}{S \sqrt{c^T(X^TX)^{-1}c}} \geq M_\alpha \right) = \alpha$

We square the expression and get  

\[
[c^T(\hat{\beta} - \beta)]^2 = [c^T(\hat{\beta} - \beta)]^T [c^T(\hat{\beta} - \beta)] = (\hat{\beta} - \beta)^T (cc^T)(\hat{\beta} - \beta)
\]

The value of the expression in $c$ is the same as in $rc$ with $r \in \mathbb{R}_0$, so we can maximise under the normalisation that $c^T(X^TX)^{-1}c = 1$.

$X^TX$ is a $p \times p$ symmetric, full rank matrix and so it can be decomposed as $X^TX = RR^T$ with $R$ a $p \times p$ invertible matrix. So, we can write that $c^T(X^TX)^{-1}c = c^T(R^{-1}R^{-1})c = u^T u$ with $u = R^{-1}c$, so $c = Ru$.

We maximise over $u$,  

\[
\max_u \frac{(\hat{\beta} - \beta)^T Ru u^T R^T (\hat{\beta} - \beta)}{u^T u}
\]

The numerator contains a squared inner product of $u$ with $R^T(\hat{\beta} - \beta)$. According to Cauchy-Schwarz's inequality:

\[
\hat{\beta} - \beta)^T Ru u^T R^T (\hat{\beta} - \beta) \leq \|u\|^2 \cdot \|R^T(\hat{\beta} - \beta)\|^2, \\
\]

So  

\[
\max_u \frac{(\hat{\beta} - \beta)^T Ru u^T R^T (\hat{\beta} - \beta)}{u^T u} = \|R^T(\hat{\beta} - \beta)\|^2 = \\
(\hat{\beta} - \beta)^T RR^T (\hat{\beta} - \beta) \sim pF_{p,n-p}
\]
**Residuals and outliers**

\[ e = Y - \hat{y} = (I - P)Y \]

\[ \Sigma_{ee} = \sigma^2 (I - P) \]

This allows to standardise/studentise the residuals and then test for normality.

**Cook-distance; influence measure; influence points**

measures the influence of observation \( i \) on the eventual estimated vector:

\[ D_i = \frac{(\hat{\beta}(i) - \hat{\beta})^T (X^T X)(\hat{\beta}(i) - \hat{\beta})}{s^2} \]

**Transformations**

**Model**

\[ Y^{(\lambda)} = X\beta + \varepsilon \]

with \( Y_i^{(\lambda)} = \frac{Y_i - \lambda}{\lambda} \)

**General: transformation of a random variable**

If \( Y = g(X) \) then

\[ f_Y(y) = f_X(x(y)) |J(y)| = f_X(g^{-1}(y)) |J(y)| \]

with \( J = \det \begin{bmatrix} \ldots & \frac{\partial x_i}{\partial y_j} & \ldots \\ \ldots & \ldots & \ldots \end{bmatrix} \) and \( \frac{\partial x_i}{\partial y_j} = \frac{\partial y_i^{-1}}{\partial y_j} \)

In our case \( Y^{(\lambda)} \sim N(X\beta, \sigma^2 I) \) takes the role of \( X \), so \( \frac{\partial x_i}{\partial y_j} = 0 \) unless \( j = i \), in that case \( \frac{\partial x_i}{\partial y_i} = y_i^{\lambda - 1} \)

So \( J = \prod_{i=1}^n y_i^{\lambda - 1} \)

**Some notions of model selection**

The analysis until now assumed that the full (global) model is correct. For instance, the expressions \( E\beta = \beta \) and \( \Sigma_{y} = \sigma^2 (X^T X)^{-1} \) are true only if indeed \( Y \) can be written as \( X\beta \) and no explanatory variable is missing.

In order to check if a specific covariate \( x_i \) should be considered as a candidate in the global model, one could "play safe" put it in the full model and let the hypothesis test \( H_{ij}: \beta_j = 0 \) decide whether or not it is significant.

However, if we have to play safe for a very wide range of possible explanatory variables, a lot of parameters need to be estimated, leaving just a few (if at all any) degrees of freedom to estimate the variance of the observational error. As a consequence, the resulting hypothesis tests will have weak power, and really significant parameters may be left undiscovered.

We would therefore prefer a procedure that limits the size of the full model before the actual estimation and testing takes place. At the same time, the model selection procedure should be able to move beyond the simple question whether or not a specific explanatory variable (or a group of covariates) should be in the model. The model selection should also be able to compare the quality of two arbitrary models where each model contains variables not present in the other.
As the philosophy of hypothesis testing favors the null hypothesis a priori (the alternative has to come up with significant data), this philosophy cannot be used in a symmetric comparison of two (or more) models.

A model selection procedure therefore makes use of a model selection criterion that expresses how well the data can be described in that model.

We define **Coefficient of determination**

\[
R^2 = \frac{SS_R}{SS_T} \quad \text{or} \quad R^2 = 1 - \frac{SS_E}{SS_T}
\]

\(R^2\) is a measure of how well the model explains the variance of the response variables. However, a submodel of a full model has always a smaller \(R^2\), even if the full model only adds unnecessary, insignificant parameters to the submodel.

**Note** The denominator in \(R^2\) is the same for all models. The numerator is (up to a constant) the log-likelihood of a normal model with known variance. So \(R^2\) measures the likelihood of a model. The more parameters the model has, the more likely the observations become, because there are more degrees of freedom to make the observations fit into the model.

**Adjusted coefficient of determination**

\[
\text{Adjusted } R^2 = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \quad \text{Or } 1 - R^2 = \frac{n-1}{n-p}(1 - R^2)
\]

**Interpretation**

We know: \(S^2 = \frac{(Y - \hat{Y})^T(Y - \hat{Y})}{n-p} = \frac{1}{n-p} \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 = \frac{1}{n-p} SS_E\)

and \(\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \hat{y}_i)^2 = \frac{n-p}{n} S^2 = \frac{1}{n} SS_E = \frac{1}{n}(SS_T - SS_R) = \frac{1}{n} SS_T(1 - R^2)\)

Suppose we have two nested models, i.e., one model containing all covariates of the other (this is: a full model with \(p\) covariates and a reduced model with \(q\) restrictions, so with \(p - q\) out of \(p\) free covariates), then the \(F\)-value in testing the significance of the full model is then:

\[
F = \frac{(\hat{\sigma}^2 - \hat{\sigma}_q^2)/(n-p)}{\hat{\sigma}_q^2/(n-p)}
\]

Dividing numerator and denominator by \(SS_T\) leads to:

\[
F = \frac{(R^2 - R_q^2)/(n-p)}{(1-R_q^2)/(n-p)} = \frac{(1-R_q^2)}{q(1-R_q^2)/(n-p)}
\]
We rewrite this in terms of “adjusted \( R^{2\text{a}} \)

\[
F = \frac{n-p}{q} \left[ \frac{1-R^2_{p-q}}{1-R^2_p} - 1 \right] + \frac{1-R^2_{p-q}}{1-R^2_p}
\]

from which it follows that \( R^2_{p-q} \leq R^2_p \Leftrightarrow F \geq 1 \)

The criterion that decides which model has higher quality is thus equivalent to a “hypothesis test” with critical value \( F = 1 \), instead of \( F_{q,n-p,q} \). This illustrates the symmetry of the model selection criterion, compared to the testing procedure, which uses null and alternative hypotheses, with clearly unequal roles.

---

The average variance can we elaborated as

\[
\text{Average Variance} = \text{var}(\hat{y}_p) = \text{var}(\hat{y}_p) + (E\hat{y}_p - y)^2
\]

with \( y = x^T \beta \) in the correct model

Because of linearity of expectations

\[
E\hat{y}_p = E(x^T_p(X^T_pX_p)^{-1}X^T_pY) = x^T_p(X^T_pX_p)^{-1}X^T_py
\]

And \( \text{var}(\hat{y}_p) = x^T_p \Sigma_p x_p = \sigma^2 x^T_p(X^T_pX_p)^{-1}x_p \)

We search for the minimum average prediction error, where the average is taken over all points of observation, so we want to minimize \( \Delta_p = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_{pi} - y_i)^2 \)

The average variance can we elaborated as

\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} \text{var}(\hat{y}_{pi}) = \sum_{i=1}^{n} x^T_{pi}(X^T_pX_p)^{-1}x_{pi}
\]

\[
= \sum_{i=1}^{n} \text{Tr} \left( x^T_{pi} (X^T_pX_p)^{-1}x_{pi} \right) = \sum_{i=1}^{n} \text{Tr} \left( (X^T_pX_p)^{-1}(x_{pi}x^T_{pi}) \right)
\]

\[
= \text{Tr} \left( I_p \right) = p
\]

thereby using \( X_p = \begin{bmatrix} x^T_{p1} \\ \vdots \end{bmatrix} \) so \( X^T_pX_p = \sum_{i=1}^{n} (x_{pi}x^T_{pi}) \)

The average variance is thus only dependent on the size of the model. It does not depend on the exact locations of the points in which the prediction quality is evaluated. We don’t need that the prediction is evaluated in the points of observations for this part.
The average mean or expected error (bias)

\[ \frac{1}{\sigma^2} \sum_{i=1}^{n} (E\hat{y}_p - y_i)^2 = \frac{1}{\sigma^2} (E\hat{y}_p - y)^T (E\hat{y}_p - y) \]

Now \( E\hat{y}_p = \left[ \begin{array}{c} x_p^T (X_p^T X_p)^{-1} X_p^T y \\ \vdots \end{array} \right] = X_p (X_p^T X_p)^{-1} X_p^T y = P_p y \)

so the average bias is \( \frac{1}{\sigma^2} y^T (I - P)^T (I - P) y \)

This contains the unknown, error-free response variable. We try to estimate the bias by plugging in the observed values instead. This is of course only possible in points of observations. This is why the information criterion uses points of observations only. We find the expression \( Y^T (I - P) Y = SS_E \).

This estimator of the bias is biased again, as follows from

\[ E(Y^T (I - P) Y) = E[(y + \varepsilon)^T (I - P)(y + \varepsilon)] \]

\[ = y^T (I - P)y + E[\varepsilon^T (I - P)y] + E[y^T (I - P) \varepsilon] + E[\varepsilon^T (I - P) \varepsilon] \]

\[ = y^T (I - P)y + 0 + 0 + \text{Tr}(I - P)\sigma^2 = y^T (I - P)y + (n - p)\sigma^2 \]

But this time, the bias, \((n - p)\sigma^2\) only depends on the model size, no longer on the unknown response.

So \( \Delta_p \sigma^2 = E[SS_E + (2p - n)\sigma^2] \)

The random variable on the right hand side (between the expectation brackets) is observable and it is an unbiased estimator of the prediction quality measure in the left hand side of the expression.

In most practical situations \( \sigma^2 \) has to be estimated by \( S^2 \) and so \( C_p = \frac{SS_E}{S^2} + 2p - n \) is a consistent estimator for \( \Delta_p \)

---

**Algorithms for searching the best model**

- Suppose the full model has \( K \) possible variables, then the number of possible models in \( 2^K \)
- Computing all adjusted \( R^2 \)'s, \( C_p \) scores or other criterion is impossible (exponential computational complexity)
- Possible strategies:
  1. **Forward selection**: Start with a zero model, then gradually add the most significant parameter in a test with \( H_0 \) the current model.
  2. **Backward elimination**
  3. **Forward-backward alternation**: after a variable has been added, it is tested if any variable in the model can now be eliminated without significant increase in RSS.
  4. **Grouped/structured selection**: the routine may impose that some parameters must be included before others. For instance, no interaction terms without main effects.
Sparsity

- **Parsimonious** model selection: we search for models where "some" covariates have exactly zero contribution.

- **Sparse** model selection: the number of zeros dominates (by far) the number of nonzeros.

Regularisation: from combinatorial to convex constraints

- In the constraint optimization problem
  \[ \text{RegLS}(\hat{\beta}) = \| Y - K \cdot \hat{\beta} \|^2 + \lambda p \]
  the constraint or penalty equals \( p = \sum_i \hat{\beta}_i^0 = \| \beta \|_0 \), where we take \( 0^0 = 0 \).

  \( \ell_0 \) regularisation = **combinatorial** integer programming problem

- Replace \( \ell_0 \) by \( \ell_1 \): \rightarrow **convex** quadratic programming problem

  \[ \text{RegLS}(\hat{\beta}) = \| Y - K \cdot \hat{\beta} \|^2 + \lambda \sum_i | \hat{\beta}_i | \]

Solving the \( \ell_1 \) constraint problem

Steepest descend for quadratic forms

Suppose \( F(\beta) = \| Y - X\beta \|^2 = \sum_{i=1}^n \left( Y_i - \sum_j X_{ij} \beta_j \right)^2 \)

then \( \frac{\partial F}{\partial \beta_k} = \sum_{i=1}^n 2 \left( Y_i - \sum_j X_{ij} \beta_j \right) X_{ik} \)

Sum over \( i \) is an inner product of the vectors \( Y - X\beta \) and the \( k \)th column of \( X \), so all \( k \) together, the gradient vector becomes \( X^T (Y - X\beta) \)

Karush-Kuhn-Tucker conditions

If \( \hat{\beta} \) solves the \( \ell_1 \) constraint optimization problem

\[ 1C_1(\beta) = \| Y - X\beta \|^2 + 2\lambda \| \beta \|_1 \] then
\[ X_j^T(Y - X\hat{\beta}) = \lambda \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0 \]
\[ |X_j^T(Y - X\hat{\beta})| \leq \lambda \text{ if } \hat{\beta}_j = 0 \]

**Intuition**

- If \( \hat{\beta}_j \neq 0 \), then \( \frac{\partial IC_1(\beta)}{\partial \hat{\beta}_j} \) is continuous, and it must be zero. That is expressed by the first line in the Karush-Kuhn-Tucker conditions.
- If \( \hat{\beta}_j = 0 \), then the partial derivative is discontinuous. The steepest decrease of \( \|Y - X\beta\|_2^2 \) by letting \( \hat{\beta}_j \) deviate from zero, should then be smaller than the increase in the penalty.

**Consequence**

The solution of the \( \ell_1 \) constraint problem will be sparse, but it does not coincide with the \( \ell_0 \) constraint solution. Indeed, the solution that satisfies the KKT conditions cannot be an orthogonal projection (a least squares solution).

On the next slides is a simple, yet important, special case: soft-versus hard-thresholding. The latter is a \( \ell_0 \), and thus a projection, the former is projection plus shrinkage.

---

**Soft- and Hard-Thresholding**

- Suppose \( X = I \), so we have the observational model \( Y = \beta + \varepsilon \). In this model, both \( \ell_0 \) and \( \ell_1 \)-penalized closeness-of-fit can be optimized componentwise.
- \( \ell_0 \)-penalized closeness-of-fit:
  \[ IC(\beta) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_i)^2 + \lambda \text{I}(|\hat{\beta}_i| > 0) \]
  Solution is **Hard-thresholding**: \( \hat{\beta}_i = \text{HT}_\lambda(Y_i) \), where \( \text{HT}_\lambda(x) \) is the following function.

- \( \ell_1 \)-penalized closeness-of-fit:
  \[ IC(\beta) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_i)^2 + 2\lambda |\hat{\beta}_i| \]
  Solution is **Soft-thresholding**: \( \hat{\beta}_i = \text{ST}_\lambda(Y_i) \), where \( \text{ST}_\lambda(x) \) is the following function.
LASSO/Basis pursuit

- $\ell_1$ penalization/regularization/constraint is called least absolute shrinkage and selection operator (LASSO) or Basis Pursuit
- For given $\lambda$, $\ell_1$ penalization leads to the same degree of sparsity as $\ell_0$ (see Figure)
- For fixed $\lambda$ and if all nonzero $\beta$ are large enough, $\ell_1$ penalization is variable selection consistent: for $n \to \infty$, the set of nonzero variables in the selection equals the true set with probability tending to one
- The convex optimization problem can be solved by quadratic/dynamic programming or by specific methods, such as
  - Least Angle Regression (LARS), a direct solver
  - Iterative Soft Thresholding (It. ST), an iterative solver

We discuss both on the subsequent slides

Least Angle Regression

- Brings in a new variable that with maximum projection of current residual, i.e., component with highest magnitude in the vector $X^T(Y - X\hat{\beta}_0)$
- Moves along equiangular vector $u_A \hat{\beta}_1 = \hat{\beta}_0 + \alpha u_A$ with $u_A = X_A(X_A^T X_A)^{-1}1_A / \sqrt{1_A^T (X_A^T X_A)^{-1}1_A}$ where $A$ is set of active variables (variables currently in the model), until one variable not in $A$ has the same projection of the new residual $X^T(Y - X\hat{\beta}_1)$
- Stopping criterion: Mallows’s $C_p$: we stop if (we think that) $C_p(p)$ has reached a minimum

Iterative soft-thresholding

- General $X$, observational model $Y = X\beta + \varepsilon$
- We want $\|Y - X\hat{\beta}\|_2^2$ as small as possible, under the constraint that $\|\hat{\beta}\|_1 = \sum |\beta_i|$ is restricted
- Suppose we have an estimator $\hat{\beta}_{(r)}$, then we improve this estimator in two steps
  - We proceed in the direction of the steepest descend in $\|Y - X\hat{\beta}\|_2^2$.
    - The steepest descend is $X^T(Y - X\hat{\beta}_{(r)})$
    - We define $\hat{\beta}_{(r+1/2)} = \hat{\beta}_{(r)} + X^T(Y - X\hat{\beta}_{(r)})$
  - We search for $\hat{\beta}_{(r+1)}$ which is as close as possible to $\hat{\beta}_{(r+1/2)}$, but which has restricted $\ell_1$-norm, so, we take $\hat{\beta}_{(r+1)} = ST_\lambda(\hat{\beta}_{(r+1/2)})$