1 Linear Model

We observe a random response variable $Y_i$ and fixed, non-random explanatory variables (predictors, covariates) $(X_{i1}, X_{i2}, \ldots, X_{ip})^T$, $i = 1, \ldots, n$ distributed according to

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p + \epsilon_i;$$

in matrix-vector notation:

$$Y = X\beta + \epsilon,$$

where $Y \in \mathbb{R}^n$ is a vector containing the $n$ response variables, $X \in \mathbb{R}^{n \times p}$ is the $n \times p$ matrix containing the explanatory variables, $\beta \in \mathbb{R}^p$ is the vector of $p$ linear parameters of the model, and $\epsilon \in \mathbb{R}^n$ is a vector of $n$ random identically-distributed measurement errors which satisfy $E(\epsilon) = 0_n$ and $\text{Cov}(\epsilon) = \sigma^2 I_n$.

Observations.

- Cov($Y$) = $\sigma^2 I_n \Rightarrow Y_i$ are uncorrelated and have same variance.
  
  $E(Y) = X\beta \Rightarrow Y_i$ are not identically distributed. They have different means according to the corresponding vector of explanatory variables $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$.

- According to the linear model, the $Y_i$'s can take values on a continuum on $\mathbb{R}$. The linear model is not an adequate description of the data if the $Y_i$'s are constrained such that they do not lie in a continuum. For instance, the $Y$'s can be integer-valued, binary, categorical, ...
  
  In the linear model the mean parameter $\mu(\beta) = E(Y) = X\beta$ is linear in the explanatory variables. For responses $Y_i$ which do not lie on a continuum, we model some function of $E(Y_i)$ as linear; e.g. for binary data $Y_i \in \{0, 1\}$ we have $\mu_i = E(Y_i) = P(Y_i = 1)$, and it makes sense to model the log-odds ratio as linear:

$$\log \left( \frac{\mu_i}{1 - \mu_i} \right) = X_i^T\beta.$$

Rearrange to obtain:

$$\mu_i(\beta) = \frac{e^{X_i^T\beta}}{1 + e^{X_i^T\beta}}.$$

To estimate $\beta$ we use maximize the likelihood of the data, e.g. maximize over $\beta$ the function:
\[ \ell(\beta) = \prod_{i=1}^{n} Y_i^{\mu_i(\beta)} (1 - Y_i)^{(1-\mu_i(\beta))}. \]

2 Least squares estimation

For a candidate estimate of \( \beta \), say \( b \in \mathbb{R}^p \), the \( i \)th residual is:

\[ Y_i - X_{i1}b_1 - X_{i2}b_2 - \ldots - X_{ip}b_p. \]

It is reasonable to choose \( b \) to minimize the residual sum of squares:

\[ S(b) = \sum_{i=1}^{n} (Y_i - X_{i1}b_1 - X_{i2}b_2 - \ldots - X_{ip}b_p)^2. \]

In matrix-vector notation, the residual sum of squares can be expressed by:

\[ S(b) = \|Y - Xb\|^2 = (Y - Xb)^T (Y - Xb), \]

where for a \( k \)-vector \( w \), \( \|w\| = w_1^2 + w_2^2 + \ldots + w_k^2 \) is the square of the usual Euclidean norm in \( k \)-dimensional space.

A value of \( b \) that minimizes \( S(b) \) is a least-squares estimate of \( \beta \). We denote a least-squares estimate of \( \beta \) by \( \hat{\beta}_{LS} \).

2.1 Characterizing \( \hat{\beta}_{LS} \).

We can write:

\[ S(b) = (Y - Xb)^T (Y - Xb) = Y^TY - 2Y^TXb + b^TX^TXb. \]

Compute the gradient vector \( \nabla S(b) = (\frac{\partial S(b)}{\partial b_1}, \ldots, \frac{\partial S(b)}{\partial b_p})^T \):

\[ \nabla S(b) = -2Y^TX + 2X^TXb. \]

Compute the Hessian matrix \( \nabla^2 S(b) \) where \( (\nabla^2 S(b))_{ij} = \frac{\partial^2 S(b)}{\partial b_i \partial b_j} \):

\[ \nabla^2 S(b) = 2X^TX. \]

Since \( X^TX \) is positive semi-definite we know \( S(b) \) is a convex function of \( b \). Therefore any \( b \) which satisfies \( \nabla S(b) = 0 \) is a global minimum of \( S \), i.e. a least-squares estimate. This yields a set of equations, called the normal equations which \( \hat{\beta}_{LS} \) must satisfy:
\((X^T X) \hat{\beta}_{LS} = Y^T X\).

Does \(\hat{\beta}_{LS}\) exist? Always:

1. If \(X^T X\) is invertible, then \(\hat{\beta}_{LS} = (X^T X)^{-1} X^T Y\) is the unique least-squares solution.

2. If \(X^T X\) is not invertible, there are an infinite number of solutions to the normal equations. We need to review some key concepts from linear algebra to prove this.

3 Linear Algebra Review

- Vectors, linear (in)dependence
- Linear subspaces
  - Basis
  - Dimension
- Orthogonality
  - Orthogonal complement of a subspace
  - Key decomposition: for any vector \(w \in \mathbb{R}^k\) and subspace \(\mathcal{V}\) we can decompose \(w = w_{\mathcal{V}} + w_{\mathcal{V}^\perp}\) where \(w_{\mathcal{V}} \in \mathcal{V}\) and \(w_{\mathcal{V}^\perp} \in \mathcal{V}^\perp\).
  * \(\dim(\mathcal{V}) + \dim(\mathcal{V}^\perp) = k\)
- Matrices
  - Column space \(\mathcal{C}(A)\) of matrix \(A\).
  - Row space \(\mathcal{R}(A) = \mathcal{C}(A^T)\)
  - \(\text{rank}(A) = \dim(\mathcal{C}(A)) = \dim(\mathcal{R}(A))\).
  - Kernel (null space) \(\mathcal{K}(A)\)
  - Result:
    \(\mathcal{K}(A) = \mathcal{R}(A)\perp\)
    \(\mathcal{K}(A^T) = \mathcal{C}(A)\perp\).