Note on Linear Models

Melody Y. Huang

1. Review of Linear Algebra

Definition 1.1 (Linear independence)
A set of vectors $\{x_1, ...x_n\}$ *is* ***linearly dependent*** *if a non-trivial combination of them is zero. More formally:*

$$\exists\ c_i \neq 0\ s.t.\ \sum_{i=1}^{n} c_i x_i = 0$$

Therefore, a set of vectors are ***linearly independent*** *if no such combination exists.*

To show a set of vectors are linearly independent (or not), we must show that $\sum_{i=1}^{n} c_i x_i = 0$ iff $c_i = 0\ \forall i$.

Example 1.1
We define the following vectors:

$$v_1 = \begin{pmatrix} 1 \\ 2 \\ 7 \end{pmatrix}, v_2 = \begin{pmatrix} 5 \\ 6 \\ 1 \end{pmatrix}, v_3 = \begin{pmatrix} 6 \\ 8 \\ 8 \end{pmatrix}$$

The set of vectors $\{v_1, v_2, v_3\}$ *are not linearly independent because* $v_3 = v_1 + v_2$.

Example 1.2
We have the following set of vectors:

$$\begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 4 \\ 2 \\ -2 \end{pmatrix}$$

We can show that these vectors are independent by solving for $a, b,$ *and* c*:*

$$a \begin{pmatrix} 2 \\ 2 \\ 0 \end{pmatrix} + b \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + c \begin{pmatrix} 4 \\ 2 \\ -2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

From solving the system of equations, we find that $a = b = c = 0$ *in order for the above to hold. Therefore, the set of vectors are linearly independent.*

Alternatively, if x_i is of n dimensions and we have n such vectors (i.e., putting all the vectors together to form a square matrix), we can compute the determinant of this matrix. If the determinant is non-zero, then the vectors are linearly independent.

Definition 1.2 (Span)
*The **span** is the set of all finite linear combinations of elements in a subspace.*

$$span\{x_1, ...x_n\} = \left\{ \sum_{i=1}^{n} c_i x_i \;\middle|\; c_1, ...c_n \in \mathbb{R} \right\}$$

Note that the linear combinations in the span of a subspace does not necessarily have to be linearly independent!

Definition 1.3 (Basis)
*The **basis** is a set of linearly independent vectors that span a subspace V.*

Elements from the basis can be used to reconstruct any element in V. We refer to the number of elements in the basis of a subspace V as the **dimension of V**. In general, we want to think about matrices as groups of vectors:

$$X = \begin{pmatrix} | & | & & | \\ x_1 & x_2 & ... & x_p \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{n \times p},$$

where:

$$x_i = \begin{pmatrix} x_{i,1} \\ \vdots \\ x_{i,n} \end{pmatrix}$$

Definition 1.4 (Image)
*The **image** or **column space** of a matrix X is the vector space spanned by the columns of X. In other words:*

$$Im(X) = span\{x_1, ...x_p\} = \left\{ \sum_{i=1}^{p} \beta_i x_i \;\middle|\; \beta_1, ...\beta_p \in \mathbb{R}^p \right\}$$

This is often denoted as $Im(\cdot)$, and is sometimes also called the range.

This is very useful in the context of linear regression (which is what we primarily care about in this course). We can think of X as our matrix of covariates. The image of X is therefore the different values that the following equation could take on (with arbitrary β values):

$$\beta_1 X_1 + \beta_2 X_2 + ...\beta_p X_p$$

The premise of linear models will be to find the optimal β values such that, given some vector Y, we can represent Y the best using these p columns of X.

Definition 1.5 (Rank)
*The **rank** of a matrix X is the number of linearly independent columns of X. It is also the dimension of the column space (or the image).*

To find the rank of a matrix: reduce matrix to row-echelon form and count the number of non-zero rows. A matrix is **full rank** if its rank is equal to the minimum of its two dimensions. More formally, if a matrix X is $n \times p$, where $p < n$, then X will be full rank if $rank(X) = p$. A matrix that is not full rank is called **rank deficient**.[1] When we are dealing with square matrices, the square matrix must be full rank in order for it to be invertible. Additionally, when a square matrix is invertible, the determinant of it will be non-zero. As such, when dealing with square matrices, we can check for whether or not it is full rank by computing the determinant of the matrix.

Definition 1.6 (Kernel)
*The **kernel** is the set of vectors that are mapped to 0 by X.*

$$ker(X) = \{\beta \in \mathbb{R}^p | X\beta = 0\}$$

To find the kernel, solve for β such that $X\beta = 0$. The Rank-Nullity Theorem states that the dimension of the image (i.e., the rank) and the dimension of the kernel (i.e., nullity) should equal the total dimension of the subspace that X maps from.

Definition 1.7 (Positive Semi-Definite)
*A matrix $X \in \mathbb{R}^{n \times n}$ is **positive semi-definite** if we can take any non-zero column vector $v \in \mathbb{R}^n$ and $v^\top X v \geq 0$.*

Example 1.3
The identity matrix is positive semi-definite.

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

We introduce some random vector z:

$$z = \begin{pmatrix} a \\ b \end{pmatrix}$$

We then see that if we compute $z^\top I z$:

$$\begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = a^2 + b^2 \geq 0$$

[1] Note: We cannot have a fuller rank matrix, which means that the rank of a matrix is naturally bound by the minimum of its two dimensions (i.e., for an arbitrary matrix A that is $m \times n$, $rank(A) \leq \min(n, m)$).

Example 1.4
Another example of a PSD matrix is the covariance matrix. To illustrate this, let $X = (X_1, X_2, ...X_p)$ *be* p *arbitrary random variables, with covariance* Σ *(where* Σ *will be* $p \times p$*). Note that the variance must be, by definition, greater than or equal to zero. As such:*

$$var(X) \geq 0$$

Introducing some arbitrary (non-random) vector a*:*

$$var(a^\top X) \geq 0$$

Expanding this out:

$$var(a^\top X) = a^\top var(X)a = a^\top \Sigma a \geq 0$$
$$\implies \Sigma \text{ is PSD.}$$

The only time in which the covariance is not positive definite is in cases in which one of the variables is an exact linear function of another. (Ask yourself: how would we check for this?)

In general, if we want to determine a matrix is positive semi-definite, we can:

1. Compute its eigenvalues. A matrix is PSD if and only if all its eigenvalues are non-negative.
2. Compute all the leading principal minors. What this means is that if we take the determinant of the upper left $k \times k$ matrix (subsetting from the main matrix, for $k = 1, ...n$), all of the determinants of these sub-matrices must be non-negative.
3. Show directly using the definition.

2. Multivariate Normal Distributions

Definition 2.1 (Multivariate Normal)
Let $y = (y_1, ..., y_n)$*.* y *is normally distributed with mean vector* $\mu = (\mu_1, ..., \mu_n)$ *and covariance matrix* Σ*. We say that* y *is multivariate normal* $N(\mu, \Sigma)$*.*

The probability density function of a multivariate normal is:

$$P(y) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^\top \Sigma^{-1}(y-\mu)\right)$$

This seems scary, but we can think of the multivariate normal distribution as a collection of individual normal distributions. In lecture, we showed that marginal distributions of $y \sim N(\mu, \Sigma)$ were

multivariate normal. We can also show the converse, which is that the joint density of a bunch of individual, normally distributed random variables can be represented as a multivariate normal distribution. I will show the simple case in which variance is constant and all the variables are independent, and leave the more complex case up to you.

Let $y_i \sim N(\mu_i, \sigma^2)$, for $i = 1, ..., n$. Note that in this set up, all the y_i's have the same variance (σ^2), but that they have different means. We will also assume, for simplicity, that they are all independent.

For a single y_i:

$$P(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$$

The nice thing is that we have assumed independence. This means that the joint density of all of the y_i's is simply the product of the individual density functions. More specifically:

$$P(y_1, ..., y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right)$$

I will denote this as $P(y)$ from now on. We can clean this up a little:

$$\begin{aligned} P(y) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\right) \end{aligned}$$

Now, note the following.

Because all the y_i's are independent, $cov(y_i, y_j) = 0$, where $i \neq j$. Therefore, the covariance matrix between all of these n y_i's will be a diagonal matrix. Additionally, we made the assumption that all the y_i's have the same variance of σ^2, so we can write the covariance matrix as a constant term σ^2, multiplied by the identity matrix:

$$cov(y) = \Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

The determinant of a diagonal matrix is going to be the product of all the diagonal entries. Therefore:

$$|\Sigma| = \left(\sigma^2\right)^n$$

Furthermore, the inverse of Σ will be simply:

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma^2} & \dots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & & \frac{1}{\sigma^2} \end{bmatrix} = \frac{1}{\sigma^2} I_n$$

Finally, we note that $\sum_{i=1}^{n}(y_i - \mu)^2 = (y-\mu)^\top(y-\mu)$.

Plugging this into our joint density function:

$$\begin{aligned} P(y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2\right) \\ &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(y-\mu)^\top(y-\mu)\right) \\ &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(y-\mu)^\top\Sigma^{-1}(y-\mu)\right) \end{aligned}$$

This is exactly the pdf of a multivariate normal, with mean vector $\mu = (\mu_1, ..., \mu_n)$!

Example 2.1
We will derive the mean and expectation of a conditional multivariate normal distribution. Let X and Y be defined as:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathrm{N}\left(0, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right).$$

We begin by defining a transformation:

$$A = \begin{bmatrix} I & 0 \\ -\Sigma_{YX}\Sigma_{XX}^{-1} & I \end{bmatrix}$$

We call:

$$\begin{bmatrix} X \\ Y' \end{bmatrix} = A \begin{bmatrix} X \\ Y \end{bmatrix}$$

As such, we notice that:

$$\begin{aligned} cov(X, Y') &= cov(X, Y - \Sigma_{YX}\Sigma_{XX}^{-1}X) \\ &= cov(X,Y) - cov(X,X)\Sigma_{XX}^{-1}\Sigma_{XY} \\ &= cov(X,Y) - \Sigma_{XX}\Sigma_{XX}^{-1}\Sigma_{XY} \\ &= 0 \end{aligned}$$

Therefore, X and Y' are independent.

$$\begin{aligned} \mathbb{E}(Y'|X) &= \mathbb{E}(Y - \Sigma_{YX}\Sigma_{XX}^{-1}X|X = x) \\ &= E(Y|X) - \Sigma_{YX}\Sigma_{XX}^{-1}x \end{aligned}$$

Because X and Y' are independent, we know that the expected value of Y', conditioned on X is simply $\mathbb{E}(Y')$. Furthermore:

$$\mathbb{E}(Y'|X) = \mathbb{E}(Y') = \mathbb{E}(Y) - \Sigma_{YX}\Sigma_{XX}^{-1}\mathbb{E}(X) = 0$$

Therefore, we can rewrite the above as:

$$\begin{aligned} E(Y|X=x) &= E(Y'|X) + \Sigma_{YX}\Sigma_{XX}^{-1}x \\ &= 0 + \Sigma_{YX}\Sigma_{XX}^{-1}x \\ &= \Sigma_{YX}\Sigma_{XX}^{-1}x \end{aligned}$$

For the variance, we once again use the fact that X and Y' are independent. This means that:

$$cov(Y'|X=x) = cov(Y')$$

We can rewrite $cov(Y'|X=x)$ as:

$$\begin{aligned} cov(Y'|X=x) &= cov(Y - \Sigma_{YX}\Sigma_{XX}^{-1}x|X=x) \\ &= cov(Y|X=x) \end{aligned}$$

$$\implies cov(Y'|X=x) = cov(Y|X=x) \tag{1}$$

We can rewrite $cov(Y')$ as:

$$cov(Y') = cov(Y - \Sigma_{YX}\Sigma_{XX}^{-1}X) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} \tag{2}$$

Setting (1) and (2) equal to each other:

$$cov(Y|X=x) = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$$

Conditional distributions of multivariate normal distributions are also normally distributed. Therefore, the conditional distribution of Y, given $X = x$ will be a normal distribution, centered at $\Sigma_{YX}\Sigma_{XX}^{-1}x$, with covariance $\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$.

3. Maximum Likelihood Estimation

Normally in statistics, we build models and want to assess how good the models are based on the actual observed data. Maximum likelihood estimation is a natural way to think about this. The main idea behind MLE is: given an assumed distribution/model, we want to pick the parameters of the model such that it maximizes the likelihood (or probability) of observing the data.

When we talk about likelihood, we're really talking about the probability density functions, conditional on some parameter value. As such, if we know (or assume) that some random variable X_i is drawn from a generic distribution f with parameter θ, then the likelihood function of X_i is the probability of observing X_i, given θ takes on some value:

$$f(X_i|\theta)$$

The likelihood function is denoted as $L(\theta|X_i)$. However, we **do not** interpret this as the probability of θ taking on certain values given X_i. It is essentially equivalent to $f(X_i|\theta)$. However, the reason

L is written as such is that we want to think of this as a function of θ and not X_i. In this context, X_i has been observed and therefore, is fixed.

$$L(\theta|X_i) = f(X_i|\theta)$$

Our goal, therefore, is to pick a θ that increases the likelihood that X_i could have been drawn.

This is much easier illustrated using an example.

Assume $Y_1, ...Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Assume for simplicity that σ^2 is known, and we are trying to find the best μ value.

The likelihood of observing a single Y_i will be:

$$P(Y_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \mu)^2\right)$$

The likelihood of observing all n values will be:

$$\begin{aligned} L(\mu|Y) = P(Y|\mu) &= \prod_{i=1}^{n} P(Y_i|\mu) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2\right) \end{aligned}$$

Our goal is to maximize the likelihood. Therefore, this is really just a simple optimization problem, where we can take the first derivative of the likelihood function and find the local optimum points. It's often times easier to work with the log-likelihood function. This is because when we have joint densities, we have a lot of products, and when we take the log of products, they become additive, which is super nice! The local optimum points found for the log-likelihood function will be the same as the optimum points found for the likelihood function. This is because the log function is strictly increasing.

$$\begin{aligned} \ell(\mu) &= \log P(Y|\mu) \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2 \end{aligned}$$

We then take the first derivative of $\ell(\mu)$ and set it equal to zero:

$$\begin{aligned} \frac{\partial}{\partial\mu}\ell(\mu) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \mu)\cdot(-1) \\ &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu) \end{aligned}$$

$$\implies \hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

In other words, $\hat{\mu}_{MLE}$ is simply the sample mean. The first derivative of the log-likelihood function is often times referred to as the **score** (denoted as $S(\cdot)$). Therefore, generally speaking, the maximum likelihood estimate is the solution to $S(\theta) = 0$.

Our example illustrates a classic MLE procedure for when we are trying to find the MLE of a single parameter. When there are multiple parameters to optimize, we can think of θ as a vector of parameters (i.e., $\theta = (\theta_1, ...\theta_p)$). The score function will be:

$$S(\theta) = \nabla \ell(\theta),$$

and as such, setting this equal to zero and solving for θ will give us the MLE of θ. However, in this case, because θ is a vector of p parameters, we will have p estimates.

4. Linear Models

When we have a set of X and we want to predict Y (some dependent variable), the best predictor of Y given X will be the conditional expectation function (i.e., $\mathbb{E}(Y|X)$). Often times, we cannot calculate the conditional expectation function directly, so we have to estimate it. If we impose some linear structure in our prediction, then the best predictor of the conditional expectation function will be the linear model of the following form:

$$Y = X\beta + \varepsilon$$

Assumptions:

- $\varepsilon|X \sim N(0, \sigma^2 I_n)$
  As such, we have assumed that the noise is centered at 0, with variance $\sigma^2 I_n$.
  This in turn implies that:
  $$\mathbb{E}(Y|X) = \mathbb{E}(X\beta + \varepsilon|X) = X\beta$$
  $$var(Y|X) = var(X\beta + \varepsilon|X) = var(\varepsilon|X) = \sigma^2 I_n$$
  Additionally, because we assume that $\varepsilon|X$ is Gaussian noise, we implicitly assume that Y is also multivariate normal, with mean $X\beta$ and variance $\sigma^2 I_n$.
- X is fixed and has full column rank.

The issue is that this relies on knowing (or being able to estimate) the true value of β, which comes from our *population*. We do not usually have the capability to estimate the population value, and

as such, need a sample analog of β. As such, we will estimate β, and from there, construct an estimate of Y (denoted $\hat{Y}$).

We denote the fitted values from this as $\hat{Y}$. The error we get from our fitted model (denoted e) will be the the difference between the actual Y value and the predicted $\hat{Y}$:

$$e = Y - \hat{Y}$$

Therefore, the optimal $\hat{\beta}$ value will be the value that *minimizes* the error in some capacity.

Usually, the error incurred by a model will correspond to some sort of penalty factor that represents how egregious the error is. This correspondence takes place through a loss function. Common loss functions are the sum of absolute error (i.e., $\sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |Y_i - \hat{Y}_i| = ||Y - \hat{Y}||_1$), or the sum of squared error ($\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = ||Y - \hat{Y}||_2^2$). These are also called ℓ_1 and ℓ_2 loss (respectively). Under ℓ_1 loss, errors that are very large are penalized by the same amount as errors that are very small. However, under ℓ_2 loss, errors that are very large are penalized further, thereby penalizing for outliers more heavily.

In our case, we will deal with sum of squared errors (ℓ_2). Therefore, our goal will be to pick a $\hat{\beta}$ to minimize the sum of squared error:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Reformulating this in more concise notation:

$$\hat{\beta} = \arg\min_{\beta} \underbrace{||Y - X\hat{\beta}||_2^2}_{e^\top e}$$

As with any minimization problem, we will take the derivative of the objective function w.r.t. to the parameter of interest (β) and set it equal to zero to solve for the optimal $\hat{\beta}$. To make our lives a little easier, we can expand out the objective function as such:

$$\begin{aligned} e^\top e &= (Y - X\hat{\beta})^\top (Y - X\hat{\beta}) \\ &= Y^\top Y - \hat{\beta}^\top X^\top Y - Y^\top X\hat{\beta} + \hat{\beta}^\top X^\top X\hat{\beta} \\ &= Y^\top Y - 2\hat{\beta}^\top X^\top Y + \hat{\beta}^\top X^\top X\hat{\beta} \end{aligned}$$

Note: $Y^\top X\hat{\beta}$ is a scalar. The transpose of a scalar is simply the scalar. As such:

$$(Y^\top X\hat{\beta})^\top = Y^\top X\hat{\beta} \implies \hat{\beta}^\top X^\top Y = Y^\top X\hat{\beta}$$

Taking the derivative with respect to $\hat{\beta}$:

$$\frac{\partial e^\top e}{\partial \hat{\beta}} = -2X^\top Y + 2X^\top X\hat{\beta} = 0$$

$$\implies \hat{\beta} = (X^\top X)^{-1} X^\top Y$$

4.1. Properties of OLS

Proposition 4.1 (Sampling Distribution of $\hat{\beta}$)
In a linear model $Y = X\beta + \varepsilon$, where $\hat{\beta}$ is given as $(X^\top X)^{-1}X^\top Y$ (i.e., $\hat{\beta}$ is the OLS estimate of β), and ε is some noise factor, the following is true:

1. *$\hat{\beta}$ is an unbiased estimate for β.*
2. *The covariance of $\hat{\beta}$ is given by $\sigma^2(X^\top X)^{-1}$*

Let's take a look at these.

1. $\hat{\beta}$ is an unbiased estimate for β.
   To begin, we rewrite $\hat{\beta}$ by making the very smart observation that we can substitute in $X\beta+\varepsilon$ for Y:

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1}X^\top Y \\ &= (X^\top X)^{-1}X^\top(X\beta+\varepsilon) \\ &= (X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top\varepsilon \\ &= \beta + (X^\top X)^{-1}X^\top\varepsilon\end{aligned}$$

   Now recall that we have assumed that $\mathbb{E}(\varepsilon|X) = 0$. As such, taking the conditional expectation of $\hat{\beta}$ given X:

$$\begin{aligned}\mathbb{E}(\hat{\beta}|X) &= \mathbb{E}(\beta + (X^\top X)^{-1}X^\top\varepsilon|X) \\ &= \beta + (X^\top X)^{-1}X^\top \underbrace{\mathbb{E}(\varepsilon|X)}_{=0} \\ &= \beta\end{aligned}$$

2. $var(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$

$$\begin{aligned}var(\hat{\beta}|X) &= var(\hat{\beta}-\beta|X) \\ &= var((X^\top X)^{-1}X^\top\varepsilon|X) \\ &= (X^\top X)^{-1}X^\top \underbrace{var(\varepsilon|X)}_{=\sigma^2}((X^\top X)^{-1}X^\top)^\top\end{aligned}$$

   By homoskedasticty, we assume $var(\varepsilon|X)$ is constant:

$$\begin{aligned}&= \sigma^2(X^\top X)^{-1}X^\top((X^\top X)^{-1}X^\top)^\top \\ &= \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1} \\ &= \sigma^2(X^\top X)^{-1}\end{aligned}$$

In reality, we do not know what σ^2 is. Therefore, we estimate it by using the sum of squared residuals:

$$\hat{\sigma}^2 = \frac{SSR}{n-p} = \frac{||e||^2}{n-p}$$

(This is also sometimes notated as s^2. Also, sum of squared residuals is often times called sum of squared errors.)

Theorem 4.1 (Gauss-Markov)
Under the following assumptions, $\hat{\beta}$ will be the best linear unbiased estimator (BLUE).[2]

1. *Linearity*
2. *Exogeneity* ($\mathbb{E}(\varepsilon_i|X) = 0$)
3. *Homoskedasticity* ($var(\varepsilon) = \sigma^2 I$; $var(\varepsilon_i) = \sigma^2$)
4. *No serial correlation* ($cov(X, \varepsilon) = 0$)
5. *No multicollinearity (equivalent to saying that X is full rank)*
6. *Optional:* $\varepsilon|X \sim N(0, \sigma^2 I)$.

Proof.
We will show this using proof by contradiction.
Let's introduce a new estimator $\tilde{\beta}$ that is defined as:

$$\tilde{\beta} = AY = \hat{\beta} + BY$$

As such:

$$\begin{aligned}
\tilde{\beta} &= ((X^\top X)^{-1}X^\top + B)Y \\
&= ((X^\top X)^{-1}X^\top + B)(X\beta + \varepsilon) \\
&= (X^\top X)^{-1}X^\top X\beta + (X^\top X)^{-1}X^\top \varepsilon + BX\beta + B\varepsilon \\
&= \beta + ((X^\top X)^{-1}X^\top + B)\varepsilon + BX\beta
\end{aligned}$$

In order for $\tilde{\beta}$ to be a valid competitor for $\hat{\beta}$, it must unbiased.

$$\mathbb{E}(\tilde{\beta}|X) = \mathbb{E}(\beta + ((X^\top X)^{-1}X^\top + B)\varepsilon + BX\beta|X)$$

[2] As a note, in this context, X (our covariates) is effectively fixed, so when we take the expectation of these quantities, we can pull X out. If you don't like the idea that all your covariates are now treated as constants, you can think of this as essentially taking the conditional expectation, given the covariates. I will explicitly write out the conditional expectation (i.e., $\mathbb{E}(e|X)$), but just know that these are equivalent.

$$= \beta + ((X^\top X)^{-1}X^\top + B)\underbrace{\mathbb{E}(\varepsilon|X)}_{=0} + BX\beta$$
$$= \beta + BX\beta$$

In order for $\mathbb{E}(\tilde{\beta}|X) = 0$, $BX = 0$. Now, for $\tilde{\beta}$ to be a better estimator than $\hat{\beta}$, it must have lower variance.

$$var(\tilde{\beta}|X) = var(\beta + ((X^\top X)^{-1}X^\top + B)\varepsilon + BX\beta)$$

Note that $BX = 0$, and β is simply a constant. Therefore, we can clean this up:

$$\begin{aligned}
&= var((X^\top X)^{-1}X^\top + B|X) \\
&= ((X^\top X)^{-1}X^\top + B)\underbrace{var(\varepsilon|X)}_{=\sigma^2 I}((X^\top X)^{-1}X^\top + B)^\top \\
&= \sigma^2((X^\top X)^{-1}X^\top + B)((X^\top X)^{-1}X^\top + B)^\top \\
&= \sigma^2((X^\top X)^{-1}X^\top + B)(X(X^\top X)^{-1} + B^\top) \\
&= \sigma^2(\underbrace{(X^\top X)^{-1}X^\top X}_{=I}(X^\top X)^{-1} + (X^\top X)^{-1}\underbrace{X^\top B^\top}_{(BX)^\top=0} + \underbrace{BX}_{=0}(X^\top X)^{-1} + BB^\top) \\
&= \sigma^2((X^\top X)^{-1} + BB^\top) \\
&= var(\hat{\beta}) + \sigma^2 BB^\top \geq var(\hat{\beta})
\end{aligned}$$

The last inequality follows from the fact that $BB^\top \geq 0$ (since we are effectively squaring each entry in B. Therefore, the variance of any alternative unbiased estimator $\tilde{\beta}$ will have worse variance than $\hat{\beta}$. As such, $\hat{\beta}$ is the best we can do for unbiased linear estimators.

The 6th (optional) Gauss-Markov assumption states that the error term is normally distributed, with mean 0 and variance $\sigma^2 I$ (i.e., $\varepsilon \sim N(0, \sigma^2 I)$). This is equivalent to imposing a distributional assumption on Y. More specifically, if we take this assumption to be true, then Y will be normally distributed with mean $X\beta$ and variance $\sigma^2 I$.

This means we can use maximum likelihood estimation to estimate β and σ^2. From before (see MLE section), we know that the likelihood function of a normally distributed random variable Y is simply:

$$\begin{aligned}
f(Y) &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(Y-\mu)^\top(Y-\mu)\right) \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}||Y-\mu||^2\right) \\
&\text{Substituting in } X\beta \text{ for } \mu\text{:} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}||Y-X\beta||^2\right)
\end{aligned}$$

The log-likelihood of this will be:

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\underbrace{||Y - X\beta||^2}_{(*)}$$

Notice that the (*) term is equivalent to the objective function in our least squares estimation problem. Taking the derivative with respect to β and setting it equal to zero:

$$\frac{\partial}{\partial\beta} = -2X^\top(Y - X\beta) = 0$$

$$\implies \hat{\beta}_{MLE} = (X^\top X)^{-1}X\top Y$$

Therefore, if Y is normally distributed, then the OLS estimate of β is also the maximum likelihood estimator.