Lecture 13: Simple Linear Regression in Matrix Format

To move beyond simple regression we need to use matrix algebra. We'll start by re-expressing simple linear regression in matrix form. Linear algebra is a pre-requisite for this class; I strongly urge you to go back to your textbook and notes for review.

1 Expectations and Variances with Vectors and Matrices

If we have p random variables, $Z_1, Z_2, \ldots Z_p$, we can put them into a random vector $\mathbf{Z} = [Z_1 Z_2 \ldots Z_p]^T$. This random vector can be thought of as a $p \times 1$ matrix of random variables.

This expected value of $\mathbf{Z}$ is defined to be the vector

$$\mu \equiv \mathbb{E}\left[Z\right] = \begin{bmatrix} \mathbb{E}\left[Z_1\right] \\ \mathbb{E}\left[Z_2\right] \\ \vdots \\ \mathbb{E}\left[Z_p\right] \end{bmatrix}. \tag{1}$$

If a and b are non-random scalars, then

$$\mathbb{E}\left[a\mathbf{Z} + b\mathbf{W}\right] = a\mathbb{E}\left[\mathbf{Z}\right] + b\mathbb{E}\left[\mathbf{W}\right]. \tag{2}$$

If $\mathbf{a}$ is a non-random vector then

$$\mathbb{E}(\mathbf{a}^T\mathbf{Z}) = \mathbf{a}^T\mathbb{E}(\mathbf{Z}).$$

If $\mathbf{A}$ is a non-random matrix, then

$$\mathbb{E}\left[\mathbf{A}\mathbf{Z}\right] = \mathbf{A}\mathbb{E}\left[\mathbf{Z}\right]. \tag{3}$$

Every coordinate of a random vector has some covariance with every other coordinate. The variance-covariance matrix of $\mathbf{Z}$ is the $p \times p$ matrix which stores these value. In other words,

$$\mathrm{Var}\left[\mathbf{Z}\right] \equiv \begin{bmatrix} \mathrm{Var}\left[Z_1\right] & \mathrm{Cov}\left[Z_1, Z_2\right] & \ldots & \mathrm{Cov}\left[Z_1, Z_p\right] \\ \mathrm{Cov}\left[Z_2, Z_1\right] & \mathrm{Var}\left[Z_2\right] & \ldots & \mathrm{Cov}\left[Z_2, Z_p\right] \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}\left[Z_p, Z_1\right] & \mathrm{Cov}\left[Z_p, Z_2\right] & \ldots & \mathrm{Var}\left[Z_p\right] \end{bmatrix}. \tag{4}$$

This inherits properties of ordinary variances and covariances. Just as $\mathrm{Var}\left[Z\right] = \mathbb{E}\left[Z^2\right] - (\mathbb{E}\left[Z\right])^2$, we have

$$\mathrm{Var}\left[\mathbf{Z}\right] = \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^T\right] - \mathbb{E}\left[\mathbf{Z}\right]\left(\mathbb{E}\left[\mathbf{Z}\right]\right)^T \tag{5}$$

For a non-random vector $\mathbf{a}$ and a non-random scalar b,

$$\mathrm{Var}\left[\mathbf{a} + b\mathbf{Z}\right] = b^2\mathrm{Var}\left[\mathbf{Z}\right]. \tag{6}$$

For a non-random matrix $\mathbf{C}$,

$$\mathrm{Var}\left[\mathbf{C}\mathbf{Z}\right] = \mathbf{C}\mathrm{Var}\left[\mathbf{Z}\right]\mathbf{C}^T. \tag{7}$$

(Check that the dimensions all conform here: if $\mathbf{c}$ is $q \times p$, $\mathrm{Var}\left[\mathbf{c}\mathbf{Z}\right]$ should be $q \times q$, and so is the right-hand side.)

A random vector $\mathbf{Z}$ has a multivariate Normal distribution with mean μ and variance Σ if its density is

$$f(\mathbf{z}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right).$$

We write this as $\mathbf{Z} \sim N(\mu, \Sigma)$ or $\mathbf{Z} \sim MVN(\mu, \Sigma)$.

If A is a square matrix, then the trace of A — denoted by A — is defined to be the sum of the diaginal elements. In other words,

$$\operatorname{tr} A = \sum_j A_{jj}.$$

Recall that the trace satisfies these properties:

$$\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B), \quad \operatorname{tr}(cA) = c\operatorname{tr}(A), \quad \operatorname{tr}(A^T) = \operatorname{tr}(A)$$

and we have the cyclic property

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB).$$

If $\mathbf{C}$ is non-random, then $\mathbf{Z}^T\mathbf{C}\mathbf{Z}$ is called a *quadratic form.* We have that

$$\mathbb{E}\left[\mathbf{Z}^T\mathbf{C}\mathbf{Z}\right] = \mathbb{E}\left[\mathbf{Z}\right]^T \mathbf{C}\mathbb{E}\left[\mathbf{Z}\right] + \operatorname{tr}[\mathbf{C}\mathrm{Var}\left[\mathbf{Z}\right]]. \tag{8}$$

To see this, notice that

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z} = \operatorname{tr} \mathbf{Z}^T\mathbf{C}\mathbf{Z} \tag{9}$$

because it's a 1×1 matrix. But the trace of a matrix product doesn't change when we cyclicly permute the matrices, so

$$\mathbf{Z}^T\mathbf{C}\mathbf{Z} = \operatorname{tr} \mathbf{C}\mathbf{Z}\mathbf{Z}^T \tag{10}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Z}^T\mathbf{C}\mathbf{Z}\right] &= \mathbb{E}\left[\operatorname{tr} \mathbf{C}\mathbf{Z}\mathbf{Z}^T\right] && (11)\\
&= \operatorname{tr} \mathbb{E}\left[\mathbf{C}\mathbf{Z}\mathbf{Z}^T\right] && (12)\\
&= \operatorname{tr} \mathbf{C}\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^T\right] && (13)\\
&= \operatorname{tr} \mathbf{C}(\mathrm{Var}\left[\mathbf{Z}\right] + \mathbb{E}\left[\mathbf{Z}\right]\mathbb{E}\left[\mathbf{Z}\right]^T) && (14)\\
&= \operatorname{tr} \mathbf{C}\mathrm{Var}\left[\mathbf{Z}\right] + \operatorname{tr} \mathbf{C}\mathbb{E}\left[\mathbf{Z}\right]\mathbb{E}\left[\mathbf{Z}\right]^T) && (15)\\
&= \operatorname{tr} \mathbf{C}\mathrm{Var}\left[\mathbf{Z}\right] + \operatorname{tr} \mathbb{E}\left[\mathbf{Z}\right]^T \mathbf{C}\mathbb{E}\left[\mathbf{Z}\right] && (16)\\
&= \operatorname{tr} \mathbf{C}\mathrm{Var}\left[\mathbf{Z}\right] + \mathbb{E}\left[\mathbf{Z}\right]^T \mathbf{C}\mathbb{E}\left[\mathbf{Z}\right] && (17)
\end{aligned}$$

using the fact that tr is a linear operation so it commutes with taking expectations; the decomposition of $\mathrm{Var}\left[\mathbf{Z}\right]$; the cyclic permutation trick again; and finally dropping tr from a scalar.

Unfortunately, there is generally no simple formula for the variance of a quadratic form, unless the random vector is Gaussian. If $\mathbf{Z} \sim N(\mu, \Sigma)$ then $\mathrm{Var}(\mathbf{Z}^T C\mathbf{Z}) = 2\operatorname{tr}(C\Sigma C\Sigma) + 4\mu^T C\Sigma C\mu$.

2 Least Squares in Matrix Form

Our data consists of n paired observations of the predictor variable X and the response variable Y, i.e., $(X_1, Y_1), \ldots (X_n, Y_n)$. We wish to fit the model

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{18}$$

where $\mathbb{E}\left[\epsilon|X = x\right] = 0$, $\mathrm{Var}\left[\epsilon|X = x\right] = \sigma^2$, and ϵ is uncorrelated across measurements.

2.1 The Basic Matrices

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}. \tag{19}$$

Note that $\mathbf{X}$ — which is called the *design matrix* — is an $n \times 2$ matrix, where the first column is always 1, and the second column contains the actual observations of X. Now

$$\mathbf{X}\beta = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix}. \tag{20}$$

So we can write the set of equations

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \ldots, n$$

in the simpler form

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

2.2 Mean Squared Error

Let

$$\mathbf{e} \equiv \mathbf{e}(\beta) = \mathbf{Y} - \mathbf{X}\beta. \tag{21}$$

The training error (or observed mean squared error) is

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n} e_i^2(\beta) = MSE(\beta) = \frac{1}{n}\mathbf{e}^T\mathbf{e}. \tag{22}$$

Let us expand this a little for further use. We have:

$$\begin{aligned} MSE(\beta) &= \frac{1}{n}\mathbf{e}^T\mathbf{e} && (23) \\ &= \frac{1}{n}(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) && (24) \\ &= \frac{1}{n}(\mathbf{Y}^T - \beta^T\mathbf{X}^T)(\mathbf{Y} - \mathbf{X}\beta) && (25) \\ &= \frac{1}{n}\left(\mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) && (26) \\ &= \frac{1}{n}\left(\mathbf{Y}^T\mathbf{Y} - 2\beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta\right) && (27) \end{aligned}$$

where we used the fact that $\beta^T\mathbf{X}^T\mathbf{Y} = (\mathbf{Y}^T\mathbf{X}\beta)^T = \mathbf{Y}^T\mathbf{X}\beta$.

2.3 Minimizing the MSE

First, we find the gradient of the MSE with respect to β:

$$\nabla MSE(\beta) = \frac{1}{n}\left(\nabla \mathbf{Y}^T\mathbf{Y} - 2\nabla \beta^T \mathbf{X}^T\mathbf{Y} + \nabla \beta^T\mathbf{X}^T\mathbf{X}\beta\right) \quad (28)$$

$$= \frac{1}{n}\left(0 - 2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\beta\right) \quad (29)$$

$$= \frac{2}{n}\left(\mathbf{X}^T\mathbf{X}\beta - \mathbf{X}^T\mathbf{Y}\right) \quad (30)$$

We now set this to zero at the optimum, $\widehat{\beta}$:

$$\mathbf{X}^T\mathbf{X}\widehat{\beta} - \mathbf{X}^T\mathbf{Y} = 0. \quad (31)$$

This equation, for the two-dimensional vector $\widehat{\beta}$, corresponds to our pair of normal or estimating equations for $\widehat{\beta}_0$ and $\widehat{\beta}_1$. Thus, it, too, is called an estimating equation. Solving, we get

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}. \quad (32)$$

That is, we've got one matrix equation which gives us both coefficient estimates.

If this is right, the equation we've got above should in fact reproduce the least-squares estimates we've already derived, which are of course

$$\widehat{\beta}_1 = \frac{c_{XY}}{s^2_X}, \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}. \quad (33)$$

Let's see if that's right.

As a first step, let's introduce normalizing factors of $1/n$ into both the matrix products:

$$\widehat{\beta} = (n^{-1}\mathbf{X}^T\mathbf{X})^{-1}(n^{-1}\mathbf{X}^T\mathbf{Y}) \quad (34)$$

Now

$$\frac{1}{n}\mathbf{X}^T\mathbf{Y} = \frac{1}{n}\begin{bmatrix} 1 & 1 & \ldots & 1 \\ X_1 & X_2 & \ldots & X_n \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad (35)$$

$$= \frac{1}{n}\begin{bmatrix} \sum_i Y_i \\ \sum_i X_i Y_i \end{bmatrix} = \begin{bmatrix} \overline{Y} \\ \overline{XY} \end{bmatrix}. \quad (36)$$

Similarly,

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \frac{1}{n}\begin{bmatrix} n & \sum_i X_i \\ \sum_i X_i & \sum_i X_i^2 \end{bmatrix} = \begin{bmatrix} 1 & \overline{X} \\ \overline{X} & \overline{X^2} \end{bmatrix}. \quad (37)$$

Hence,

$$\left(\frac{1}{n}\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{\overline{X^2} - \overline{X}^2}\begin{bmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix} = \frac{1}{s^2_X}\begin{bmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}.$$

Therefore,

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} &= \frac{1}{s_X^2}\begin{bmatrix} \overline{X^2} & -\overline{X} \\ -\overline{X} & 1 \end{bmatrix}\begin{bmatrix} \overline{Y} \\ \overline{XY} \end{bmatrix} && (38)\\
&= \frac{1}{s_X^2}\begin{bmatrix} \overline{X^2}\,\overline{Y} - \overline{X}\,\overline{XY} \\ -(\overline{X}\;\overline{Y}) + \overline{XY} \end{bmatrix} && (39)\\
&= \frac{1}{s_X^2}\begin{bmatrix} (s_X^2 + \overline{X}^2)\overline{Y} - \overline{X}(c_{XY} + \overline{X}\;\overline{Y}) \\ c_{XY} \end{bmatrix} && (40)\\
&= \frac{1}{s_X^2}\begin{bmatrix} s_x^2\overline{Y} + \overline{X}^2\overline{Y} - \overline{X}c_{XY} - \overline{X}^2\overline{Y} \\ c_{XY} \end{bmatrix} && (41)\\
&= \begin{bmatrix} \overline{Y} - \frac{c_{XY}}{s_X^2}\overline{X} \\ \frac{c_{XY}}{s_X^2} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}. && (42)
\end{aligned}
$$

3 Fitted Values and Residuals

Remember that when the coefficient vector is β, the point predictions (fitted values) for each data point are $\mathbf{X}\beta$. Thus the vector of fitted values is

$$\widehat{\mathbf{Y}} \equiv \widehat{\mathbf{m}(\mathbf{X})} \equiv \widehat{\mathbf{m}} = \mathbf{X}\widehat{\beta}.$$

Using our equation for $\widehat{\beta}$, we then have

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

where

$$\mathbf{H} \equiv \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{43}$$

is called the *hat matrix* or the *influence matrix.*

Let's look at some of the properties of the hat matrix.

1. *Influence.* Check that $\partial\widehat{\mathbf{Y}}_i/\partial Y_j = H_{ij}$. Thus, H_{ij} is the rate at which the i^{th} fitted value changes as we vary the j^{th} observation, the "influence" that observation has on that fitted value.

2. *Symmetry.* It's easy to see that $\mathbf{H}^T = \mathbf{H}$.

3. *Idempotency.* Check that $\mathbf{H}^2 = \mathbf{H}$, so the matrix is idempotent.

Geometry. A symmtric, idempotent matrix is a projection matrix. This means that $\mathbf{H}$ projects $\mathbf{Y}$ into a lower dimensional subspace. Specifically, $\mathbf{Y}$ is a point in $\mathbb{R}^n$ but $\widehat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$ is a linear combination of two vectors, namely, the two columns of $\mathbf{X}$. In other words:

$\mathbf{H}$ projects $\mathbf{Y}$ onto the column space of $\mathbf{X}$.

The column space of $\mathbf{X}$ is the set of vectors that can be written as linear combinations of the columns of $\mathbf{X}$.

3.1 Residuals

The vector of residuals, $\mathbf{e}$, is

$$\mathbf{e} \equiv \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{44}$$

Here are some properties of $\mathbf{I} - \mathbf{H}$:

1. *Influence.* $\partial e_i / \partial y_j = (\mathbf{I} - \mathbf{H})_{ij}$.
2. *Symmetry.* $(\mathbf{I} - \mathbf{H})^T = \mathbf{I} - \mathbf{H}$.
3. *Idempotency.* $(\mathbf{I} - \mathbf{H})^2 = (\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}^2$. But, since $\mathbf{H}$ is idempotent, $\mathbf{H}^2 = \mathbf{H}$, and thus $(\mathbf{I} - \mathbf{H})^2 = (\mathbf{I} - \mathbf{H})$.

Thus,

$$MSE(\widehat{\beta}) = \frac{1}{n}\mathbf{Y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{Y} = \frac{1}{n}\mathbf{Y}^T(\mathbf{I} - \mathbf{H})\mathbf{Y}. \tag{45}$$

3.2 Expectations and Covariances

Remember that $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ where ϵ is an $n \times 1$ matrix of random variables, with mean vector $\mathbf{0}$, and variance-covariance matrix $\sigma^2\mathbf{I}$. What can we deduce from this?

First, the expectation of the fitted values:

$$\begin{aligned} \mathbb{E}[\widehat{\mathbf{Y}}] &= \mathbb{E}\left[\mathbf{HY}\right] = \mathbf{H}\mathbb{E}\left[\mathbf{Y}\right] = \mathbf{HX}\beta + \mathbf{H}\mathbb{E}\left[\epsilon\right] & (46) \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + 0 = \mathbf{X}\beta. & (47) \end{aligned}$$

Next, the variance-covariance of the fitted values:

$$\begin{aligned} \mathrm{Var}\left[\widehat{\mathbf{Y}}\right] &= \mathrm{Var}\left[\mathbf{HY}\right] = \mathrm{Var}\left[\mathbf{H}(\mathbf{X}\beta + \epsilon)\right] & (48) \\ &= \mathrm{Var}\left[\mathbf{H}\epsilon\right] = \mathbf{H}\mathrm{Var}\left[\epsilon\right]\mathbf{H}^T = \sigma^2\mathbf{HIH} = \sigma^2\mathbf{H} & (49) \end{aligned}$$

using the symmetry and idempotency of $\mathbf{H}$.

Similarly, the expected residual vector is zero:

$$\mathbb{E}\left[\mathbf{e}\right] = (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbb{E}\left[\epsilon\right]) = \mathbf{X}\beta - \mathbf{X}\beta = 0. \tag{50}$$

The variance-covariance matrix of the residuals:

$$\begin{aligned} \mathrm{Var}\left[\mathbf{e}\right] &= \mathrm{Var}\left[(\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \epsilon)\right] & (51) \\ &= \mathrm{Var}\left[(\mathbf{I} - \mathbf{H})\epsilon\right] & (52) \\ &= (\mathbf{I} - \mathbf{H})\mathrm{Var}\left[\epsilon\right](\mathbf{I} - \mathbf{H}))^T & (53) \\ &= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T & (54) \\ &= \sigma^2(\mathbf{I} - \mathbf{H}) & (55) \end{aligned}$$

Thus, the variance of each residual is not quite σ^2, nor are the residuals exactly uncorrelated.

Finally, the expected MSE is

$$\mathbb{E}\left[\frac{1}{n}\mathbf{e}^T\mathbf{e}\right] = \frac{1}{n}\mathbb{E}\left[\epsilon^T(\mathbf{I} - \mathbf{H})\epsilon\right]. \tag{56}$$

We know that this must be $(n-2)\sigma^2/n$.

4 Sampling Distribution of Estimators

Let's now assume that $\epsilon_i \sim N(0, \sigma^2)$, and are independent of each other and of X. The vector of all n noise terms, ϵ, is an $n \times 1$ matrix. Its distribution is a **multivariate Gaussian** or **multivariate Normal** with mean vector $\mathbf{0}$, and variance-covariance matrix $\sigma^2\mathbf{I}$. We write this as $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. We may use this to get the sampling distribution of the estimator $\widehat{\beta}$:

$$\begin{aligned}
\widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} && (57)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\beta + \epsilon) && (58)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon && (59)\\
&= \beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon && (60)
\end{aligned}$$

Since ϵ is Gaussian and is being multiplied by a non-random matrix, $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$ is also Gaussian. Its mean vector is

$$\mathbb{E}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}\left[\epsilon\right] = \mathbf{0} \qquad (61)$$

while its variance matrix is

$$\begin{aligned}
\mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}\left[\epsilon\right]\left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\right)^T && (62)\\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (63)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} && (64)\\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} && (65)
\end{aligned}$$

Since $\mathrm{Var}\left[\widehat{\beta}\right] = \mathrm{Var}\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon\right]$, we conclude that that

$$\widehat{\beta} \sim N(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}). \qquad (66)$$

Re-writing slightly,

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{n}(n^{-1}\mathbf{X}^T\mathbf{X})^{-1}\right) \qquad (67)$$

will make it easier to prove to yourself that, according to this, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are both unbiased, that $\mathrm{Var}\left[\widehat{\beta}_1\right] = \frac{\sigma^2}{n}s_X^2$, and that $\mathrm{Var}\left[\widehat{\beta}_0\right] = \frac{\sigma^2}{n}(1 + \overline{X}^2/s_X^2)$. This will also give us $\mathrm{Cov}\left[\widehat{\beta}_0, \widehat{\beta}_1\right]$, which otherwise would be tedious to calculate.

I will leave you to show, in a similar way, that the fitted values $\mathbf{HY}$ are multivariate Gaussian, as are the residuals $\mathbf{e}$, and to find both their mean vectors and their variance matrices.

5 Derivatives with Respect to Vectors

This is a brief review of basic vector calculus.

Consider some scalar function of a vector, say $f(\mathbf{X})$, where $\mathbf{X}$ is represented as a $p \times 1$ matrix. (Here $\mathbf{X}$ is just being used as a place-holder or generic variable; it's not necessarily the design matrix of a regression.) We would like to think about the derivatives of f with respect to $\mathbf{X}$. We can write $f(\mathbf{X}) = f(x_1, \ldots, x_p)$ where $\mathbf{X} = (x_1, \ldots, x_p)^T$.

The gradient of f is the vector of partial derivatives:

$$\nabla f \equiv \left[\begin{array}{c} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{array} \right]. \tag{68}$$

The first order Taylor series of f around $\mathbf{X}^0$ is

$$\begin{aligned} f(\mathbf{X}) &\approx f(\mathbf{X}^0) + \sum_{i=1}^{p} (\mathbf{X} - \mathbf{X}^0)_i \left. \frac{\partial f}{\partial x_i} \right|_{\mathbf{X}^0} && (69) \\ &= f(\mathbf{X}^0) + (\mathbf{X} - \mathbf{X}^0)^T \nabla f(\mathbf{X}^0). && (70) \end{aligned}$$

Here are some properties of the gradient:

1. *Linearity.*

$$\nabla \left(af(\mathbf{X}) + bg(\mathbf{X}) \right) = a\nabla f(\mathbf{X}) + b\nabla g(\mathbf{X}) \tag{71}$$

   PROOF: Directly from the linearity of partial derivatives.

2. *Linear forms.* If $f(\mathbf{X}) = \mathbf{X}^T\mathbf{a}$, with $\mathbf{a}$ not a function of $\mathbf{X}$, then

$$\nabla(\mathbf{X}^T\mathbf{a}) = \mathbf{a} \tag{72}$$

   PROOF: $f(\mathbf{X}) = \sum_i X_i a_i$, so $\partial f/\partial X_i = a_i$. Notice that $\mathbf{a}$ was already a $p \times 1$ matrix, so we don't have to transpose anything to get the derivative.

3. *Linear forms the other way.* If $f(\mathbf{X}) = \mathbf{b}\mathbf{X}$, with $\mathbf{b}$ not a function of $\mathbf{X}$, then

$$\nabla(\mathbf{b}\mathbf{X}) = \mathbf{b}^T \tag{73}$$

   PROOF: Once again, $\partial f/\partial X_i = b_i$, but now remember that $\mathbf{b}$ was a $1 \times p$ matrix, and ∇f is $p \times 1$, so we need to transpose.

4. *Quadratic forms.* Let $\mathbf{C}$ be a $p \times p$ matrix which is not a function of $\mathbf{X}$, and consider the **quadratic form** $\mathbf{X}^T\mathbf{C}\mathbf{X}$. (You can check that this is scalar.) The gradient is

$$\nabla(\mathbf{X}^T\mathbf{C}\mathbf{X}) = (\mathbf{C} + \mathbf{C}^T)\mathbf{X}. \tag{74}$$

   PROOF: First, write out the matrix multiplications as explicit sums:

$$\mathbf{X}^T\mathbf{C}\mathbf{X} = \sum_{j=1}^{p} x_j \sum_{k=1}^{p} c_{jk} x_k = \sum_{j=1}^{p} \sum_{k=1}^{p} x_j c_{jk} x_k. \tag{75}$$

   Now take the derivative with respect to x_i:

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{p} \sum_{k=1}^{p} \frac{\partial x_j c_{jk} x_k}{\partial X_i} \tag{76}$$

If $j = k = i$, the term in the inner sum is $2c_{ii}x_i$. If $j = i$ but $k \neq i$, the term in the inner sum is $c_{ik}x_k$. If $j \neq i$ but $k = i$, we get $x_j c_{ji}$. Finally, if $j \neq i$ and $k \neq i$, we get zero. The $j = i$ terms add up to $(\mathbf{cX})_i$. The $k = i$ terms add up to $(\mathbf{c}^T\mathbf{X})_i$. (This splits the $2c_{ii}x_i$ evenly between them.) Thus,

$$\frac{\partial f}{\partial x_i} = ((\mathbf{c} + \mathbf{c}^T\mathbf{X})_i \tag{77}$$

and

$$\nabla f = (\mathbf{c} + \mathbf{c}^T)\mathbf{X}. \tag{78}$$

(Check that this has the right dimensions.)

5. *Symmetric quadratic forms.* If $\mathbf{c} = \mathbf{c}^T$, then

$$\nabla \mathbf{X}^T\mathbf{c}\mathbf{X} = 2\mathbf{c}\mathbf{X}. \tag{79}$$

5.1 Second Derivatives

The $p \times p$ matrix of second partial derivatives is called the **Hessian**. I won't step through its properties, except to note that they, too, follow from the basic rules for partial derivatives.

5.2 Maxima and Minima

We need all the partial derivatives to be equal to zero at a minimum or maximum. This means that the gradient must be zero there. At a minimum, the Hessian must be positive-definite (so that moves away from the minimum always increase the function); at a maximum, the Hessian must be negative definite (so moves away always decrease the function). If the Hessian is neither positive nor negative definite, the point is neither a minimum nor a maximum, but a "saddle" (since moving in some directions increases the function but moving in others decreases it, as though one were at the center of a horse's saddle).