11:55 Wednesday 14th October, 2015
See updates and corrections at http://www.stat.cmu.edu/~cshalizi/mreg/

Lecture 13: Simple Linear Regression in Matrix Format

36-401, Section B, Fall 2015

13 October 2015

Contents

So far, we have not used any notions, or notation, that goes beyond basic algebra and calculus (and probability). This has forced us to do a fair amount of book-keeping, as it were by hand. This is just about tolerable for the simple linear model, with one predictor variable. It will get intolerable if we have multiple predictor variables. Fortunately, a little application of linear algebra will let us abstract away from a lot of the book-keeping details, and make multiple linear regression hardly more complicated than the simple version[1].

These notes will not remind you of how matrix algebra works. However, they will review some results about *calculus* with matrices, and about expectations and variances with vectors and matrices.

Throughout, bold-faced letters will denote matrices, as $\mathbf{a}$ as opposed to a scalar a.

1 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

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

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

1.1 The Basic Matrices

Group all of the observations of the response into a single column $(n \times 1)$ matrix $\mathbf{y}$,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{2}$$

Similarly, we group both the coefficients into a single vector (i.e., a 2×1 matrix)

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{3}$$

We'd also like to group the observations of the predictor variable together, but we need something which looks a little unusual at first:

$$\mathbf{x} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{4}$$

[1] Historically, linear models with multiple predictors evolved before the use of matrix algebra for regression. You may imagine the resulting drudgery.

[2] When I need to also assume that ϵ is Gaussian, and strengthen "uncorrelated" to "independent", I'll say so.

This is an $n \times 2$ matrix, where the first column is always 1, and the second column contains the actual observations of X. We have this apparently redundant first column because of what it does for us when we multiply $\mathbf{x}$ by β:

$$\mathbf{x}\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} \tag{5}$$

That is, $\mathbf{x}\beta$ is the $n \times 1$ matrix which contains the point predictions.

The matrix $\mathbf{x}$ is sometimes called the **design matrix**.

1.2 Mean Squared Error

At each data point, using the coefficients β results in some error of prediction, so we have n prediction errors. These form a vector:

$$\mathbf{e}(\beta) = \mathbf{y} - \mathbf{x}\beta \tag{6}$$

(You can check that this subtracts an $n \times 1$ matrix from an $n \times 1$ matrix.)

When we derived the least squares estimator, we used the mean squared error,

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n} e_i^2(\beta) \tag{7}$$

How might we express this in terms of our matrices? I claim that the correct form is

$$MSE(\beta) = \frac{1}{n}\mathbf{e}^T\mathbf{e} \tag{8}$$

To see this, look at what the matrix multiplication really involves:

$$[e_1 e_2 \dots e_n] \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{9}$$

This, clearly equals $\sum_i e_i^2$, so the MSE has the claimed form.

Let us expand this a little for further use.

$$\begin{aligned} MSE(\beta) &= \frac{1}{n}\mathbf{e}^T\mathbf{e} && (10)\\ &= \frac{1}{n}(\mathbf{y} - \mathbf{x}\beta)^T(\mathbf{y} - \mathbf{x}\beta) && (11)\\ &= \frac{1}{n}(\mathbf{y}^T - \beta^T\mathbf{x}^T)(\mathbf{y} - \mathbf{x}\beta) && (12)\\ &= \frac{1}{n}\left(\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{x}\beta - \beta^T\mathbf{x}^T\mathbf{y} + \beta^T\mathbf{x}^T\mathbf{x}\beta\right) && (13) \end{aligned}$$

Notice that $(\mathbf{y}^T\mathbf{x}\beta)^T = \beta^T\mathbf{x}^T\mathbf{y}$. Further notice that this is a 1×1 matrix, so $\mathbf{y}^T\mathbf{x}\beta = \beta^T\mathbf{x}^T\mathbf{y}$. Thus

$$MSE(\beta) = \frac{1}{n}\left(\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{x}^T\mathbf{y} + \beta^T\mathbf{x}^T\mathbf{x}\beta\right) \tag{14}$$

1.3 Minimizing the MSE

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

$$\begin{aligned} \nabla MSE(\beta &= \frac{1}{n}\left(\nabla\mathbf{y}^T\mathbf{y} - 2\nabla\beta^T\mathbf{x}^T\mathbf{y} + \nabla\beta^T\mathbf{x}^T\mathbf{x}\beta\right) && (15)\\ &= \frac{1}{n}\left(0 - 2\mathbf{x}^T\mathbf{y} + 2\mathbf{x}^T\mathbf{x}\beta\right) && (16)\\ &= \frac{2}{n}\left(\mathbf{x}^T\mathbf{x}\beta - \mathbf{x}^T\mathbf{y}\right) && (17) \end{aligned}$$

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

$$\mathbf{x}^T\mathbf{x}\widehat{\beta} - \mathbf{x}^T\mathbf{y} = 0 \tag{18}$$

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

Solving,

$$\widehat{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{19}$$

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

$$\hat{\beta}_1 = \frac{c_{XY}}{s_X^2} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \tag{20}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{21}$$

Let's see if that's right.

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

$$\widehat{\beta} = (n^{-1}\mathbf{x}^T\mathbf{x})^{-1}(n^{-1}\mathbf{x}^T\mathbf{y}) \tag{22}$$

Now let's look at the two factors in parentheses separately, from right to left.

$$\begin{aligned} \frac{1}{n}\mathbf{x}^T\mathbf{y} &= \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} && (23)\\ &= \frac{1}{n}\begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} && (24)\\ &= \begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} && (25) \end{aligned}$$

Similarly for the other factor:

$$\frac{1}{n}\mathbf{x}^T\mathbf{x} = \frac{1}{n}\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \tag{26}$$

$$= \begin{bmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{bmatrix} \tag{27}$$

Now we need to take the inverse:

$$\left(\frac{1}{n}\mathbf{x}^T\mathbf{x}\right)^{-1} = \frac{1}{\overline{x^2} - \overline{x}^2}\begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \tag{28}$$

$$= \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix} \tag{29}$$

Let's multiply together the pieces.

$$(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} = \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{bmatrix}\begin{bmatrix} \overline{y} \\ \overline{xy} \end{bmatrix} \tag{30}$$

$$= \frac{1}{s_X^2}\begin{bmatrix} \overline{x^2}\overline{y} - \overline{x}\,\overline{xy} \\ -\overline{x}\overline{y} + \overline{xy} \end{bmatrix} \tag{31}$$

$$= \frac{1}{s_X^2}\begin{bmatrix} (s_X^2 + \bar{x}^2)\overline{y} - \overline{x}(c_{XY} + \bar{x}\bar{y}) \\ c_{XY} \end{bmatrix} \tag{32}$$

$$= \frac{1}{s_X^2}\begin{bmatrix} s_x^2\overline{y} + \bar{x}^2\overline{y} - \overline{x}c_{XY} - \overline{x}^2\bar{y} \\ c_{XY} \end{bmatrix} \tag{33}$$

$$= \begin{bmatrix} \overline{y} - \frac{c_{XY}}{s_X^2}\overline{x} \\ \frac{c_{XY}}{s_X^2} \end{bmatrix} \tag{34}$$

which is what it should be.

So: $n^{-1}\mathbf{x}^T\mathbf{y}$ is keeping track of $\overline{y}$ and $\overline{xy}$, and $n^{-1}\mathbf{x}^T\mathbf{x}$ keeps track of $\overline{x}$ and $\overline{x^2}$. The matrix inversion and multiplication then handles all the book-keeping to put these pieces together to get the appropriate (sample) variances, covariance, and intercepts. We don't have to remember that any more; we can just remember the one matrix equation, and then trust the linear algebra to take care of the details.

2 Fitted Values and Residuals

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

$$\widehat{\mathbf{m}} = \mathbf{x}\widehat{\beta} \tag{35}$$

Using our equation for $\widehat{\beta}$,

$$\widehat{\mathbf{m}} = \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y} \tag{36}$$

Notice that the fitted values are linear in $\mathbf{y}$. The matrix

$$\mathbf{H} \equiv \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T \tag{37}$$

does not depend on $\mathbf{y}$ at all, but does control the fitted values:

$$\widehat{\mathbf{m}} = \mathbf{H}\mathbf{y} \tag{38}$$

If we repeat our experiment (survey, observation...) many times at the same $\mathbf{x}$, we get different $\mathbf{y}$ every time. But $\mathbf{H}$ does not change. The properties of the fitted values are thus largely determined by the properties of $\mathbf{H}$. It thus deserves a name; it's usually called the **hat matrix**, for obvious reasons, or, if we want to sound more respectable, the **influence matrix**.

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

1. *Influence* Since $\mathbf{H}$ is not a function of $\mathbf{y}$, we can easily verify that $\partial \widehat{m}_i / \partial y_j = H_{ij}$. 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* A square matrix $\mathbf{a}$ is called **idempotent**[3] when $\mathbf{a}^2 = \mathbf{a}$ (and so $\mathbf{a}^k = \mathbf{a}$ for any higher power k). Again, by writing out the multiplication, $\mathbf{H}^2 = \mathbf{H}$, so it's idempotent.

Idemopotency, Projection, Geometry Idempotency seems like the most obscure of these properties, but it's actually one of the more important. $\mathbf{y}$ and $\widehat{\mathbf{m}}$ are n-dimensional vectors. If we project a vector $\mathbf{u}$ on to the line in the direction of the length-one vector $\mathbf{v}$, we get

$$\mathbf{v}\mathbf{v}^T\mathbf{u} \tag{39}$$

(Check the dimensions: $\mathbf{u}$ and $\mathbf{v}$ are both $n \times 1$, so $\mathbf{v}^T$ is $1 \times n$, and $\mathbf{v}^T\mathbf{u}$ is 1×1.) If we group the first two terms together, like so,

$$(\mathbf{v}\mathbf{v}^T)\mathbf{u} \tag{40}$$

where $\mathbf{v}\mathbf{v}^T$ is the $n \times n$ **project matrix** or **projection operator** for that line. Since $\mathbf{v}$ is a unit vector, $\mathbf{v}^T\mathbf{v} = 1$, and

$$(\mathbf{v}\mathbf{v}^T)(\mathbf{v}\mathbf{v}^T) = \mathbf{v}\mathbf{v}^T \tag{41}$$

so the projection operator for the line is idempotent. The geometric meaning of idempotency here is that once we've projected $\mathbf{u}$ on to the line, projecting its image on to the same line doesn't change anything.

Extending this same reasoning, for any linear subspace of the n-dimensional space, there is always some $n \times n$ matrix which projects vectors in arbitrary

[3] From the Latin *idem*, "same", and *potens*, "power".

position down into the subspace, and this projection matrix is always idempotent. It is a bit more convoluted to prove that any idempotent matrix is the projection matrix for some subspace, but that's also true. We will see later how to read off the dimension of the subspace from the properties of its projection matrix.

2.1 Residuals

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

$$\mathbf{e} \equiv \mathbf{y} - \mathbf{x}\widehat{\beta} \tag{42}$$

Using the hat matrix,

$$\mathbf{e} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{43}$$

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,

$$MSE(\widehat{\beta}) = \frac{1}{n}\mathbf{y}^T(\mathbf{I} - \mathbf{H})^T(\mathbf{I} - \mathbf{H})\mathbf{y} \tag{44}$$

simplifies to

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

2.2 Expectations and Covariances

We can of course consider the vector of random variables $\mathbf{Y}$. By our modeling assumptions,

$$\mathbf{Y} = \mathbf{x}\beta + \epsilon \tag{46}$$

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:

$$\begin{aligned}
\mathbb{E}[\mathbf{H}\mathbf{Y}] &= \mathbf{H}\mathbb{E}[\mathbf{Y}] && (47)\\
&= \mathbf{H}\mathbf{x}\beta + \mathbf{H}\mathbb{E}[\epsilon] && (48)\\
&= \mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{x}\beta + 0 && (49)\\
&= \mathbf{x}\beta && (50)
\end{aligned}$$

which is as it should be, since the fitted values are unbiased.

Next, the variance-covariance of the fitted values:

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

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

Similarly, the expected residual vector is zero:

$$\mathbb{E}\left[\mathbf{e}\right] = (\mathbf{I} - \mathbf{H})(\mathbf{x}\beta + \mathbb{E}\left[\epsilon\right]) = \mathbf{x}\beta - \mathbf{x}\beta = 0 \tag{56}$$

The variance-covariance matrix of the residuals:

$$\begin{aligned}
\mathrm{Var}\left[\mathbf{e}\right] &= \mathrm{Var}\left[(\mathbf{I} - \mathbf{H})(\mathbf{x}\beta + \epsilon)\right] && (57)\\
&= \mathrm{Var}\left[(\mathbf{I} - \mathbf{H})\epsilon\right] && (58)\\
&= (\mathbf{I} - \mathbf{H})\mathrm{Var}\left[\epsilon\right](\mathbf{I} - \mathbf{H}))^T && (59)\\
&= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})^T && (60)\\
&= \sigma^2(\mathbf{I} - \mathbf{H}) && (61)
\end{aligned}$$

Thus, the variance of each residual is not quite σ^2, nor (unless $\mathbf{H}$ is diagonal) are the residuals exactly uncorrelated with each other.

Finally, the expected MSE is

$$\mathbb{E}\left[\frac{1}{n}\mathbf{e}^T\mathbf{e}\right] \tag{62}$$

which is

$$\frac{1}{n}\mathbb{E}\left[\epsilon^T(\mathbf{I} - \mathbf{H})\epsilon\right] \tag{63}$$

We know (because we proved it in the exam) that this must be $(n-2)\sigma^2/n$; we'll see next time how to show this.

3 Sampling Distribution of Estimators

Let's now "turn on" the Gaussian-noise assumption, so the noise terms ϵ_i all have the distribution $N(0, \sigma^2)$, 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**[4], with mean vector $\mathbf{0}$, and

[4]Some people write this distribution as MVN, and others also as N. I will stick to the former.

variance-covariance matrix $\sigma^2\mathbf{I}$. We may use this to get the sampling distribution of the estimator $\widehat{\beta}$:

$$\begin{aligned}
\widehat{\beta} &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{Y} && (64)\\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T(\mathbf{x}\beta + \epsilon) && (65)\\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{x}\beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon && (66)\\
&= \beta + (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon && (67)
\end{aligned}$$

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

$$\mathbb{E}\left[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon\right] = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbb{E}\left[\epsilon\right] = \mathbf{0} \tag{68}$$

while its variance matrix is

$$\begin{aligned}
\mathrm{Var}\left[(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\epsilon\right] &= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathrm{Var}\left[\epsilon\right]\left((\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\right)^T && (69)\\
&= (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\sigma^2\mathbf{I}\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} && (70)\\
&= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-1} && (71)\\
&= \sigma^2(\mathbf{x}^T\mathbf{x})^{-1} && (72)
\end{aligned}$$

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

$$\widehat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}) \tag{73}$$

Re-writing slightly,

$$\widehat{\beta} \sim MVN(\beta, \frac{\sigma^2}{n}(n^{-1}\mathbf{x}^T\mathbf{x})^{-1}) \tag{74}$$

will make it easier to prove to yourself that, according to this, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are both unbiased (which we know is right), that $\mathrm{Var}\left[\hat{\beta}_1\right] = \frac{\sigma^2}{n}s_X^2$ (which we know is right) and that $\mathrm{Var}\left[\hat{\beta}_0\right] = \frac{\sigma^2}{n}(1 + \bar{x}^2/s_X^2)$ (which we know is right). This will also give us $\mathrm{Cov}\left[\hat{beta}_0, \hat{\beta}_1\right]$, 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.

4 Derivatives with Respect to Vectors

This is a brief review, not intended as a full course in 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 $\mathbf{x}$.

Derivatives are rates of change; they tell us how rapidly the function changes in response to minute changes in its arguments. Since $\mathbf{x}$ is a $p \times 1$ matrix, we could also write

$$f(\mathbf{x}) = f(x_1, x_1, x_p) \tag{75}$$

This makes it clear that f will have a partial derivative with respect to each component of $\mathbf{x}$. How much does f change when we vary the components? We can find this out by using a Taylor expansion. If we pick some base value of the matrix $\mathbf{x}^0$ and expand around it,

$$\begin{aligned} f(\mathbf{x}) &\approx f(\mathbf{x}^0) + \sum_{i=1}^{p} (\mathbf{x} - \mathbf{x}^0)_i \left.\frac{\partial f}{\partial x_i}\right|_{\mathbf{x}^0} & (76)\\ &= f(\mathbf{x}^0) + (\mathbf{x} - \mathbf{x}^0)^T \nabla f(\mathbf{x}^0) & (77) \end{aligned}$$

where we *define* the gradient, ∇f, to be the vector of partial derivatives,

$$\nabla f \equiv \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_p} \end{bmatrix} \tag{78}$$

Notice that this defines ∇f to be a one-column matrix, just as $\mathbf{x}$ was taken to be. You may sometimes encounter people who want it to be a one-*row* matrix; it comes to the same thing, but you may have to track a lot of transposes to make use of their math.

All of the properties of the gradient can be proved using those of partial derivatives. Here are some basic ones we'll need.

1. *Linearity*
$$\nabla \left(af(\mathbf{x}) + bg(\mathbf{x})\right) = a\nabla f(\mathbf{x}) + b\nabla g(\mathbf{x}) \tag{79}$$
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
$$\nabla(\mathbf{x}^T\mathbf{a}) = \mathbf{a} \tag{80}$$
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
$$\nabla(\mathbf{b}\mathbf{x}) = \mathbf{b}^T \tag{81}$$
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
$$\nabla(\mathbf{x}^T\mathbf{c}\mathbf{x}) = (\mathbf{c} + \mathbf{c}^T)\mathbf{x} \tag{82}$$
PROOF: First, write out the matrix multiplications as explicit sums:
$$\mathbf{x}^T\mathbf{c}\mathbf{x} = \sum_{j=1}^{p} x_j \sum_{k=1}^{p} c_{jk}x_k = \sum_{j=1}^{p}\sum_{k=1}^{p} x_j c_{jk} x_k \tag{83}$$
Now take the derivative with respect to x_i.
$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^{p}\sum_{k=1}^{p} \frac{\partial x_j c_{jk} x_k}{\partial x_i} \tag{84}$$
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 $(\mathbf{c}\mathbf{x})_i$. The $k = i$ terms add up to $(\mathbf{c}^T\mathbf{x})_i$. (This splits the $2c_{ii}x_i$ evenly between them.) Thus
$$\frac{\partial f}{\partial x_i} = ((\mathbf{c} + \mathbf{c}^T\mathbf{x})_i \tag{85}$$
and
$$\nabla f = (\mathbf{c} + \mathbf{c}^T)\mathbf{x} \tag{86}$$
(You can, and should, double check that this has the right dimensions.)

5. *Symmetric quadratic forms* If $\mathbf{c} = \mathbf{c}^T$, then
$$\nabla \mathbf{x}^T\mathbf{c}\mathbf{x} = 2\mathbf{c}\mathbf{x} \tag{87}$$

4.1 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.

4.2 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).

5 Expectations and Variances with Vectors and Matrices

If we have p random variables, $Z_1, Z_2, \ldots Z_p$, we can grow them into a random vector $\mathbf{Z} = [Z_1 Z_2 \ldots Z_p]^T$. (That is, the random vector is an $n \times 1$ matrix of random variables.)

This has an expected value:

$$\mathbb{E}[Z] \equiv \int \mathbf{z} p(\mathbf{z}) d\mathbf{z} \tag{88}$$

and a little thought shows

$$\mathbb{E}[Z] = \begin{bmatrix} \mathbb{E}[Z_1] \\ \mathbb{E}[Z_2] \\ \vdots \\ \mathbb{E}[Z_p] \end{bmatrix} \tag{89}$$

Since expectations of random scalars are linear, so are expectations of random vectors: when a and b are non-random scalars,

$$\mathbb{E}[a\mathbf{Z} + b\mathbf{W}] = a\mathbb{E}[\mathbf{Z}] + b\mathbb{E}[\mathbf{W}] \tag{90}$$

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

$$\mathbb{E}[\mathbf{a}\mathbf{Z}] = \mathbf{a}\mathbb{E}[\mathbf{Z}] \tag{91}$$

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:

$$\mathrm{Var}[\mathbf{Z}] \equiv \begin{bmatrix} \mathrm{Var}[Z_1] & \mathrm{Cov}[Z_1, Z_2] & \ldots & \mathrm{Cov}[Z_1, Z_p] \\ \mathrm{Cov}[Z_2, Z_1] & \mathrm{Var}[Z_2] & \ldots & \mathrm{Cov}[Z_2, Z_p] \\ \vdots & \vdots & \vdots & \vdots \\ \mathrm{Cov}[Z_p, Z_1] & \mathrm{Cov}[Z_p, Z_2] & \ldots & \mathrm{Var}[Z_p] \end{bmatrix} \tag{92}$$

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

$$\mathrm{Var}[\mathbf{Z}] = \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^T\right] - \mathbb{E}[\mathbf{Z}](\mathbb{E}[\mathbf{Z}])^T \tag{93}$$

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

$$\mathrm{Var}[\mathbf{a} + b\mathbf{Z}] = b^2 \mathrm{Var}[\mathbf{Z}] \tag{94}$$

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

$$\mathrm{Var}[\mathbf{c}\mathbf{Z}] = \mathbf{c}\mathrm{Var}[\mathbf{Z}]\mathbf{c}^T \tag{95}$$

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

For a quadratic form, $\mathbf{Z}^T\mathbf{cZ}$, with non-random $\mathbf{c}$, the expectation value is

$$\mathbb{E}\left[\mathbf{Z}^T\mathbf{cZ}\right] = \mathbb{E}[\mathbf{Z}]^T \mathbf{c}\mathbb{E}[\mathbf{Z}] + \mathrm{tr}\,\mathbf{c}\mathrm{Var}[\mathbf{Z}] \tag{96}$$

where tr is of course the trace of a matrix, the sum of its diagonal entries. To see this, notice that

$$\mathbf{Z}^T\mathbf{cZ} = \mathrm{tr}\,\mathbf{Z}^T\mathbf{cZ} \tag{97}$$

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

$$\mathbf{Z}^T\mathbf{cZ} = \mathrm{tr}\,\mathbf{cZZ}^T \tag{98}$$

Therefore

$$\begin{aligned}
\mathbb{E}\left[\mathbf{Z}^T\mathbf{cZ}\right] &= \mathbb{E}\left[\mathrm{tr}\,\mathbf{cZZ}^T\right] && (99)\\
&= \mathrm{tr}\,\mathbb{E}\left[\mathbf{cZZ}^T\right] && (100)\\
&= \mathrm{tr}\,\mathbf{c}\mathbb{E}\left[\mathbf{ZZ}^T\right] && (101)\\
&= \mathrm{tr}\,\mathbf{c}(\mathrm{Var}[\mathbf{Z}] + \mathbb{E}[\mathbf{Z}]\,\mathbb{E}[\mathbf{Z}]^T) && (102)\\
&= \mathrm{tr}\,\mathbf{c}\mathrm{Var}[\mathbf{Z}] + \mathrm{tr}\,\mathbf{c}\mathbb{E}[\mathbf{Z}]\,\mathbb{E}[\mathbf{Z}]^T) && (103)\\
&= \mathrm{tr}\,\mathbf{c}\mathrm{Var}[\mathbf{Z}] + \mathrm{tr}\,\mathbb{E}[\mathbf{Z}]^T\,\mathbf{c}\mathbb{E}[\mathbf{Z}] && (104)\\
&= \mathrm{tr}\,\mathbf{c}\mathrm{Var}[\mathbf{Z}] + \mathbb{E}[\mathbf{Z}]^T\,\mathbf{c}\mathbb{E}[\mathbf{Z}] && (105)
\end{aligned}$$

using the fact that tr is a linear operation so it commutes with taking expectations; the decomposition of $\mathrm{Var}[\mathbf{Z}]$; 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.

6 Further Reading

Linear algebra is a pre-requisite for this class; I strongly urge you to go back to your textbook and notes for review, if any of this is rusty. If you desire additional resources, I recommend Axler (1996) as a concise but thorough course. Petersen and Pedersen (2012), while not an introduction or even really a review, is an extremely handy compendium of matrix, and matrix-calculus, identities.

References

Axler, Sheldon (1996). *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Berlin: Springer-Verlag.

Petersen, Kaare Brandt and Michael Syskind Pedersen (2012). *The Matrix Cookbook*. Tech. rep., Technical University of Denmark, Intelligent Signal Processing Group. URL `http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=3274`.