I am assuming that most of you have taken a course in linear algebra, and that the rest of you have had enough experience with vectors, dot products, and matrices that you can assimilate this small part of linear algebra quickly. It would be a good idea to have a linear algebra text to refer to during this course. What appears in this first section is not intended as an introduction to linear algebra but rather as a distillation of those parts of linear algebra most needed in linear models. The treatment here is brief, but you will be expected to have a command of it. Please consult me any time about any part of this writeup you do not understand.

We write vectors in $\mathbb{R}^n$ as column matrices $[x_1, \ldots, x_n]^T$. This enables us to represent homogeneous linear functions (linear functions with no additive constants) $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$ as left multiplication by $m \times n$ matrices $A$: we send $x$ to $Ax$. So whenever we have an $m \times n$ matrix $A$ we will understand that we can think of it as the linear function $x \rightarrow Ax$.

With this understanding the range of the function $A$ is the linear span of the columns of $A$.

Writing vectors as column matrices also enables us to represent the dot product as matrix multiplication: $x \cdot y = x^T y = \Sigma_i x_i y_i$. You will recall that the length of a vector $x$ in $\mathbb{R}^n$ is $\|x\| = (\Sigma_i x_i^2)^{1/2} = x^T x^{1/2}$. Furthermore, two vectors $x, y$ in $\mathbb{R}^n$ are orthogonal (i.e., perpendicular) if and only if $x^T y = 0$, and more generally we have $x^T y = \|x\| \|y\| \cos \theta$, where $\theta$ is the angle between $x$ and $y$.

For an $n \times n$ matrix $A$ the inverse $A^{-1}$ can be obtained whenever it exists by row reduction: first form the $n \times (2n)$ matrix $[A \ I]$. Then use row reductions to change the $A$ part of this $n \times (2n)$ matrix to $I$; in the meantime $I$ will have changed to $A^{-1}$.

To project the single vector $y$ onto the linear span of the independent vectors $x_1, \ldots, x_n$, form the matrix $X$ whose columns are the vectors $x_1, \ldots, x_n$; then the projection matrix is $P = X(X^T X)^{-1} X^T$. The desired projection of $y$ is then $Py$. To project the single vector $y$ onto the span of the single vector $x$ is simpler: the desired projection is

$$\frac{y^T x}{x^T x} x.$$

We can also avoid the matrix formula for $P$ when we want the projection $Py$ of $y$ onto the linear span of vectors $x_1, \ldots, x_n$ which are orthogonal to one another: $Py$ is then the
sum of the (easily obtained) projections of \( y \) onto the one dimensional subspaces spanned by \( x_1, \ldots, x_n \) in turn. We emphasize that this shortcut requires orthogonality of the \( x_i \)’s.

**Special matrices.** An \( n \times n \) matrix \( P \) is a projection matrix if both \( P^T = P \) and \( P^2 = P \). Geometrically, \( Py \) is the projection of \( y \) onto the column space of \( P \) for all vectors \( y \). Furthermore, the projection matrices \( P \) and \( Q \) have orthogonal ranges if and only if \( PQ = 0 \).

The \( n \times n \) matrix \( R \) is **orthogonal** if the columns of \( R \) are orthogonal and all have length 1 (are orthonormal.) Equivalently, \( R \) is orthogonal if the rows of \( R \) are orthonormal. Equivalently if \( R^T = R^{-1} \). Geometrically, the orthogonal matrices are those which preserve inner products (i.e., \((Rx)^T Ry = x^T y\)) or in other words, preserve both lengths and angles. Examples of orthogonal matrices viewed as linear functions from Euclidean space to itself are rotations in any dimension, reflections across a line through the origin in \( \mathbb{R}^2 \), and reflections across a plane through the origin in \( \mathbb{R}^3 \).

### II. Least squares and overdetermined systems.

Suppose we seek a model \( y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \) where \( y \) is the “response variable” and \( x_1, \ldots, x_k \) are the “predictors” or “explanatory variables” or “covariates”. The point is to determine \( \beta_0, \ldots, \beta_k \) from data on \( N \) cases relating \( y \) to \( x_1, \ldots, x_k \), where \( N \) is much larger than \( k \). In other words, we seek to solve

\[
y_1 = \beta_0 + \beta_1 (x_1)_1 + \ldots + \beta_k (x_k)_1
\]

\[
y_N = \beta_0 + \beta_1 (x_1)_N + \ldots + \beta_k (x_k)_N
\]

(2.1)

for \( \beta_0, \ldots, \beta_k \), the \( x \)’s and the \( y \)’s being known. Since we have many more equations than unknowns we can’t expect an exact solution very often. So we want the best compromise solution possible. If for each \( i \) we let

\[
\hat{y}_i = \beta_0 + \beta_1 (x_1)_i + \ldots + \beta_k (x_k)_i
\]

(2.2)

and if each case should contribute equally to the determination of the solution, then one highly successful method is that of least squares: i.e., choose \( \beta_0, \beta_1, \ldots, \beta_k \) to minimize

\[
\Sigma_i (y_i - \hat{y}_i)^2
\]

(2.3)
We can reformulate this setup in vector-matrix-geometric terms as follows: let
\[ y = [y_1 \ldots y_N]^T, \quad \hat{y} = [\hat{y}_1 \ldots \hat{y}_N]^T, \quad x_i = [(x_i)_1 \ldots (x_i)_N]^T \] for each \( i \), \( \hat{\beta} = [\beta_0 \ldots \beta_k]^T \), and \( 1 = [1 \ldots 1]^T \). Let \( X \) be the \( N \times (k + 1) \) matrix whose first column is \( 1 \) and whose remaining columns are \( x_1, \ldots, x_k \). Then the equations (2.1) can be expressed
\[ y = X\hat{\beta}. \tag{2.4} \]
As we said, this usually has no exact solution. The definition (2.2) can be expressed
\[ \hat{y} = X\hat{\beta} \tag{2.5}, \]
and the sum of squares (2.3) can then be expressed \( \|y - \hat{y}\|^2 \). Geometrically the possible \( \hat{y} \)'s are the possible \( X\hat{\beta} \)'s; i.e., they make up the column space of \( X \). To minimize (2.3), therefore, we need to choose \( \hat{\beta} \) so that \( \hat{y} \) is the projection of \( y \) onto the column space of \( X \); i.e.,
\[ \hat{y} = X(X^TX)^{-1}X^Ty. \tag{2.6} \]
Now since our estimate \( \hat{\beta} \) should be determined by \( \hat{y} = X\hat{\beta} \), we replace \( \hat{y} \) in (2.6) by \( X\hat{\beta} \) and arrive after some matrix algebra at
\[ \hat{\beta} = (X^TX)^{-1}X^Ty. \tag{2.7} \]

**III. Random vectors and key multivariate distributions**

A random vector \( X = [X_1 \ldots X_n]^T \) is a vector whose coordinates are one dimensional random variables. The mean vector \( E(X) \) is the vector whose coordinates are the means of the corresponding coordinates of \( X \). The covariance matrix \( \Sigma(X) = E((X - \bar{X})(X - \bar{X})^T) \) is the matrix whose \( ij \)th entry is \( \text{Cov}(X_i, X_j) \). In particular the \( i \)th diagonal element of \( \Sigma(X) \) is the variance of \( X_i \). The covariance matrix is nonnegative-definite, or in most garden variety situations positive-definite.

We have these rules, for random vector \( X \), constant matrix \( A \), and constant vector \( b \):
\[
0. \quad E(AX + b) = AE(X) + b \\
00. \quad \Sigma(AX + b) = A\Sigma(X)A^T
\]

The random vector \( X \) has the **multivariate normal distribution** with mean vector \( \bar{X} \) and covariance matrix \( \Sigma \) if the density of \( X \) over \( \mathbb{R}^n \) is given by
\[
f(x) = (\det(2\pi\Sigma))^{-n/2} \exp(-1/2[(x - \bar{X})^T\Sigma^{-1}(x - \bar{X})])
\]
for all vectors \( x \) in \( \mathbb{R}^n \). In this case we write \( X \sim N(\vec{\mu}, \Sigma) \).

We have these rules:

1. If \( X \sim N(\vec{\mu}, \Sigma) \) and \( A \) is a constant \( k \times n \) matrix whose column space is all of \( \mathbb{R}^k \) and \( b \) is a constant \( k \)-vector, then \( AX + b \sim N(A\vec{\mu} + b, A\Sigma A^T) \).

2. If \( X \sim N(\vec{\mu}, \Sigma) \) and \( A \) and \( B \) are constant matrices then \( AX \) and \( BX \) are independent if and only if \( A\Sigma B^T = 0 \).

3. \( X \sim N(\mu_1, \sigma^2 I) \) if and only if \( X_1, \ldots, X_n \) are independent one dimensional normal random variables, each with mean \( \mu \) and variance \( \sigma^2 \). In this case, if we set \( Z = (X - \mu_1)/\sigma \), then \( Z \sim N(0, I) \).

4. If \( Z \sim N(0, I) \), and \( R \) is an orthogonal matrix, then also \( RZ \sim N(0, I) \).

5. If \( Z \sim N(0, I) \) and \( P \) is a projection matrix, then \( \|PZ\|^2 \) has the chi square distribution with degrees of freedom equal to the rank of \( P \), i.e., to the dimension of the range of \( P \). Therefore if \( X \sim N(\vec{\mu}, \sigma^2 I) \), then \( \|P[(X - \vec{\mu})/\sigma]\|^2 \) also has the chi square distribution with degrees of freedom equal to the rank of \( P \).

6. If \( U \) and \( V \) are independent chi square random variables with \( m \) and \( n \) degrees of freedom respectively, then \( U + V \) is chi square with \( m + n \) degrees of freedom and \( (U/m)/(V/n) \) has the \( F \) distribution with \( m \) and \( n \) degrees of freedom.

7. If \( Z \) is a one dimensional normal random variable with mean 0 and variance 1 and \( U \) is a chi square random variable independent of \( Z \) and with \( k \) degrees of freedom, then

\[
t = \frac{Z}{(U/k)^{1/2}}
\]

has Student’s t distribution with \( k \) degrees of freedom.

8. If \( X \sim N(\vec{\mu}, \sigma^2 I) \), and \( P \) is a projection matrix, then \( E(\|PX\|^2) = \sigma^2 \text{rank}(P) + \|P(\vec{\mu})\|^2 \).

**IV. Applications to Ideal Linear Regression.**

The classic "fixed effects" regression model, in which the unaccounted for portion of the response variable is regarded as “noise”, is

\[
y = X\vec{\beta} + \vec{\epsilon},
\]

where \( y \) is an \( N \times 1 \) vector of observed responses, \( X \) is the \( N \times (k + 1) \) matrix of known ("fixed") covariates, \( \vec{\beta} \) is the \( (k + 1) \times 1 \) vector of unknown parameters, about which we wish to draw conclusions, and \( \vec{\epsilon} \) is the \( N \times 1 \) vector of “error” terms. Recall that usually our model allows for an additive constant term \( \beta_0 \) and in that case the first column of \( X \) is \( 1 \). From section II, we have \( \hat{y} = X(X^TX)^{-1}X^Ty \) and \( \hat{\beta} = (X^TX)^{-1}X^Ty \). We let
\( \mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} \) be the “hat” matrix. This is the projection matrix onto the column space of \( \mathbf{X} \) and is called the hat matrix because it puts the “hat” on \( \mathbf{y} \). We notice also that \( (\mathbf{I} - \mathbf{H})\mathbf{y} = \hat{\mathbf{e}} \) is the vector of estimated residuals. We have \( \mathbf{y} = \hat{\mathbf{y}} + \hat{\mathbf{e}} \), where \( \hat{\mathbf{y}} \) and \( \hat{\mathbf{e}} \) are perpendicular. We have observed in class that the projection of \( \mathbf{y} \) on \( \mathbf{1} \) is \( \bar{\mathbf{y}}\mathbf{1} \). Since \( \hat{\mathbf{y}} \) is the projection of \( \mathbf{y} \) onto the column space of \( \mathbf{X} \) and \( \mathbf{1} \) is in the column space of \( \mathbf{X} \), \( \bar{\mathbf{y}}\mathbf{1} \) is also the projection of \( \hat{\mathbf{y}} \) onto \( \mathbf{1} \). So we have a perpendicular decomposition:

\[
\mathbf{y} - \bar{\mathbf{y}}\mathbf{1} = (\hat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}) + \hat{\mathbf{e}}. \tag{4.2}
\]

By the Pythagorean theorem, it follows that

\[
\|\mathbf{y} - \bar{\mathbf{y}}\mathbf{1}\|^2 = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\mathbf{1}\|^2 + \|\hat{\mathbf{e}}\|^2. \tag{4.3}
\]

Written out in coordinates this is the analysis of variance for linear regression:

\[
\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2. \tag{4.4}
\]

The three sums are known as the total sum of squares \( SSY \), the regression sum of squares \( SS_{reg} \), and the residual sum of squares \( RSS \), respectively. They measure the total variation in the variable \( y \), the portion of that variation accounted for by the model, and the portion of that variation unaccounted for by the model. The ratio of the regression sum of squares to the total sum of squares can be shown in the case of one explanatory variable to be the square of the sample correlation between the explanatory and response variables. This generalizes to the case of many explanatory variables; this ratio is then the square of the multiple correlation between the response variable and the explanatory variables. This appears on printouts as “\( R^2 \)”, and is a measure of the extent to which the model accounts for the variation in the response variable.

Now we impose the standard assumption that the ”noise” vector

\[
\hat{\mathbf{e}} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4.5}
\]

where \( \sigma^2 \) is an unknown positive scalar. We sketch the main consequences of this assumption for inference: we can obtain the distributions of the estimator \( \hat{\beta} \) and its components; we can write down an estimator \( \hat{\sigma}^2 \) for \( \sigma^2 \) with distribution and as a result carry out tests on and confidence intervals for \( \beta \) and its components; and we can test whether all the covariates in a proposed model carry significant explanatory power, or whether, at least for the data before us, a particular subset of the covariates do just as good a job in accounting for the variation in the response variable.
We proceed to the particulars: from (4.1) and (4.5) we have
\[ y \sim N(X\bar{\beta}, \sigma^2 I). \] (4.6)

From (2.7) and rule 1 in section III we get
\[ \hat{\beta} \sim N(\bar{\beta}, \sigma^2 (X^T X)^{-1}). \] (4.7)

Thus the components
\[ \hat{\beta}_i \sim N(\beta_i, \sigma^2 (X^T X)^{-1}_{ii}), \] (4.8)

and we also have
\[ \text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 (X^T X)^{-1}_{ij}. \] (4.9)

We now recall that the hat matrix \( H \) is the projection onto the column space of \( X \) and of course that \( \hat{y} = Hy \). From section III, rules 5 and 2 we get that \( \|\hat{y}\|^2 / \sigma^2 \) and \( \|y - \hat{y}\|^2 / \sigma^2 \) are independent random variables, and that the latter always has the chi-squared distribution with \( N - \text{rank}(X) \) degrees of freedom. Therefore \( RSS = \|y - \hat{y}\|^2 \) has expected value \((N - \text{rank}(X))\sigma^2\), and so we take as our estimator of \( \sigma^2 \)
\[ \hat{\sigma}^2 = RSS / (N - \text{rank}(X)). \] (4.10)

Now \( \hat{\beta} \) and \( \hat{\sigma}^2 \) are independent. According to (4.8) we have for each \( i \) that \( \hat{\beta}_i \sim N(\beta_i, \sigma^2 (X^T X)^{-1}_{ii}) \), while \( [N - \text{rank}(X)]\hat{\sigma}^2 \) is chi-squared with \( N - \text{rank}(X) \) degrees of freedom. It follows from section III, rule 7 that
\[ \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}}} \] (4.11)
has Student’s t distribution with \( N - \text{rank}(X) \) degrees of freedom. The various estimators mentioned in (4.11) appear in computer printouts from standard regression packages: \( \hat{\beta}_i \) is the estimate of the coefficient \( \beta_i \), \( \hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}} \) is the standard error or standard deviation of that estimator, and the ratio of these two is the ”t-ratio” since it is the statistic used to test \( \beta_i = 0 \). Note that this ratio is not appropriate for testing whether \( \beta_i \) is equal to some specified nonzero constant. Finally note that
\[ \hat{\beta}_i \pm t_{\alpha/2} \hat{\sigma}\sqrt{(X^T X)^{-1}_{ii}} \] (4.12)
provides a \( 100(1 - \alpha) \) percent confidence interval formula for \( \beta_i \).
Last we address the questions of whether our proposed model has significant explanatory power at all, and if so, whether some specified subset of the covariates would do just as well in explaining the variation in the response variable, at least for the data at hand. Usually we are not interested in whether the constant term $\beta_0 = 0$, and so we make a preliminary adjustment. We let $P_0$ be the projection matrix onto the one dimensional subspace spanned by the vector $1$; then we have

$$P_0 y = \bar{y} 1,$$  \hspace{1cm} (4.13).

To go beyond the descriptive statistic $R^2$ to reach "firm" conclusions about the adequacy of our models, we must continue to assume (4.5). Again, $RSS/\sigma^2$ is always chi-squared with $N - \text{rank}(X)$ degrees of freedom. Furthermore from rules 5 and 8 of section III $SS_{reg}/\sigma^2$ is chi-squared if $\beta_1 = \beta_2 = \ldots = \beta_k = 0$, and is inflated otherwise (i.e., is inflated if any coefficient other than $\beta_0$ is nonzero.) Moreover these sums of squares are independent. Therefore

$$F = \frac{SS_{reg}/k}{RSS/(N - \text{rank}(X))}$$  \hspace{1cm} (4.14)

has the F distribution with $k$ and $N - k - 1$ degrees of freedom if $\beta_1 = \ldots = \beta_k = 0$ and is inflated otherwise. Therefore if this F statistic is too large according to an F table we have evidence that $\beta_i \neq 0$ for at least one $i \geq 1$; we would interpret this to mean that our model has significant explanatory power.

Next we want to ask whether given a "restricted" model $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ and a "full" model $y = \beta_0 + \ldots + \beta_p x_p + \ldots + \beta_k x_k$, does the full model have significantly more explanatory power than the restricted model, base on the data at hand? We begin with an analysis of variance

$$SS_Y = SS_{\text{reg}}^{\text{estr}} + (SS_{\text{reg}}^{\text{full}} - SS_{\text{reg}}^{\text{estr}}) + RSS^{\text{full}}.$$  \hspace{1cm} (4.15)

This analysis of variance is based on the hat matrix $H$ corresponding to the full model and the restricted hat matrix $H^{\text{estr}}$ arising from the design matrix $X^{\text{estr}}$ whose columns are the covariates in the restricted model. Thus the range of $H^{\text{estr}}$ is a subspace of the column space of $X$ generally including the vector $1$. We define $SS_{\text{reg}}^{\text{estr}} = \|(H^{\text{estr}} - P_0)y\|^2$ and $SS_{\text{reg}}^{\text{full}} = SS_{\text{reg}} = \|(H - P_0)y\|^2$. Now (4.15) follows at once. We note that this decomposition into three terms is perpendicular and independent, and that the degrees of freedom associated with the three terms are $p$, $k - p$, and $N - k - 1$ respectively. Also, let us define $RSS^{\text{estr}} = \|(I - H^{\text{estr}})y\|^2$; we already have $RSS = \|(I - H)y\|^2$. Then for computation we may note that the middle term (4.15) is equal to $RSS^{\text{estr}} - RSS^{\text{full}}$, and
this last term may be more easily obtainable from your printouts. The same considerations as before lead us to the statistic

\[ F = \frac{(SS_{\text{full}}^{\text{reg}} - SS_{\text{restr}}^{\text{reg}})/(k - p)}{RSS_{\text{full}}/(N - k - 1)}. \]  

(4.16)

This will have the F distribution with \( k - p \) and \( N - k - 1 \) degrees of freedom if \( \beta_{p+1} = \ldots = \beta_k = 0 \), and will be inflated otherwise. So if this F statistic is too large, according to the F table, we have evidence that the full model has significant additional explanatory power beyond that of the restricted model.

**V. Weighted linear regression.**

A mild generalization of the model (4.1) with assumption (4.5) is the model

\[ y = X\beta + \epsilon, \]  

(5.1)

, where now instead of (4.5) we assume

\[ \epsilon \sim N(0, \sigma^2 \Sigma), \]  

(5.2)

where \( \Sigma \) is any known positive definite matrix. An important special case is that of \( \Sigma = D \), where \( D \) is a diagonal matrix with positive entries. Let \( C \) be any invertible matrix with \( C^T C = \Sigma^{-1} \), so that \( \Sigma = C^{-1} (C^T)^{-1} \). Then if we multiply (5.1) on the left by \( C \) we get

\[ z = M\beta + \delta, \]  

(5.3)

where \( z = Cy, \ M = CX \), and \( \delta = C\epsilon \). Furthermore

\[ \delta \sim N(0, \sigma^2 I). \]  

(5.4)

We can now apply the theory of (4.1) to (5.3) and translate back. In particular for (5.3) we get

\[ \hat{\beta} = (M^T M)^{-1} M^T z = [X^T \Sigma^{-1} X]^{-1} X^T \Sigma^{-1} y. \]  

(5.5)

**VI. Mathematics behind case deletion.**

We want to determine how the key formulas in least square estimation change if we delete the \( i \)th case. To keep the notation simple we will assume \( i = 1 \); this will have no effect on the validity of the arguments.

We take our cue from the well-known formula for the sum of a geometric series:

\[ (1 - r)^{-1} = 1 + r + r^2 + \ldots, \quad |r| < 1. \]
We have also for a matrix $A$ that

$$(I - A)^{-1} = I + A + A^2 + \ldots$$

whenever we can make sense of the right hand side. Write the original design matrix $X$ in block form, with $x_1^T$ denoting the first row and corresponding to the first case, and $X_{(1)}$ denoting the design matrix from the remaining $N - 1$ cases and occupying the remaining rows of $X$. Then it is not hard to see that $X^T X = x_1 x_1^T + X_{(1)}^T X_{(1)}$. It follows that

$$X_{(1)}^T X_{(1)} = X^T X - x_1 x_1^T = [I - x_1 x_1^T (X^T X)^{-1}](X^T X).$$

Now set $A = x_1 x_1^T (X^T X)^{-1}$, and note that $A^2 = h_{11} A$ and for every $n$, $A^n = h_{11}^{n-1} A$. It follows that

$$(X_{(1)}^T X_{(1)})^{-1} = (X^T X)^{-1} [I + A(1 + h_{11} + h_{11}^2 + \ldots)] = (X^T X)^{-1} [I + \frac{1}{1-h_{11}} A] = (X^T X)^{-1} + \frac{1}{1-h_{11}} (X^T X)^{-1} x_1 x_1^T (X^T X)^{-1},$$

which is formula (5A.1) in our text.

Next we consider the effect on $\hat{\beta}$ of case deletion. We first observe that $X^T \tilde{y} = y_1 x_1 + X_{(1)}^T \tilde{y}_{(1)}$, so that $X_{(1)}^T \tilde{y}_{(1)} = X^T \tilde{y} - y_1 x_1$. It follows that

$$\hat{\beta}_{(1)} = (X_{(1)}^T X_{(1)})^{-1} X_{(1)}^T \tilde{y}_{(1)}$$

$$= [(X^T X)^{-1} + \frac{1}{1-h_{11}} (X^T X)^{-1} x_1 x_1^T (X^T X)^{-1}](X^T \tilde{y} - y_1 x_1)$$

$$= \ldots = \hat{\beta} + \frac{1}{1-h_{11}} [x_1^T \hat{\beta} - y_1][(X^T X)^{-1} x_1].$$

As a result we get

$$\hat{\beta}_{(1)} = \hat{\beta} - \frac{\hat{e}_1 (X^T X)^{-1} x_1}{1-h_{11}}.$$

Let us turn now to the residual sum of squares and other matters.