A More Careful Development of the Linear Regression Model

1 Estimating the Coefficients in the Multiple Regression Model

The *multiple regression* model that we have been working with may be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$. As usual, this stands for

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n
\end{aligned}$$

or, in matrix notation,

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \quad \underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_{n\times n})$$

i.e.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

We know that

$$\begin{aligned}
\text{Mean: } E[y_i] &= E[\beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \epsilon_i] \\
&= \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + E[\epsilon_i] \\
&= \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + 0 \\
&= \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik}, \text{ and}
\end{aligned}$$

$$\begin{aligned}
\text{Variance: } \mathrm{Var}\,(y_i) &= \mathrm{Var}\,(\beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \epsilon_i) \\
&= \mathrm{Var}\,(\epsilon_i) \\
&= \sigma^2.
\end{aligned}$$

which we write compactly in matrix notation as

$$\begin{aligned}
E[\underline{y}] &= X\underline{\beta} \\
\mathrm{Var}\,(\underline{y}) &= \sigma^2 I_{n\times n}, \text{ and so} \\
\underline{Y} &\sim N(X\underline{\beta},\ \sigma^2 I_{n\times n})
\end{aligned}$$

If we have some guesses

$$\underline{\hat{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$$

for $\underline{\beta}$, then the fitted values are

$$\underline{\hat{y}} = X\underline{\hat{\beta}}$$

and the residuals are

$$\underline{E} = \underline{y} - \underline{\hat{y}}$$

To find good estimates of $\underline{\beta}$, we may minimize the sum of squares of residuals,

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k) &= \sum_{i=1}^n E_i^2 \;=\; \sum_{i=1}^n (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots \hat{\beta}_k x_{ik}])^2 \\ &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= (\underline{y} - \underline{\hat{y}})^T (\underline{y} - \underline{\hat{y}}) \\ &= (\underline{y} - X\underline{\hat{\beta}})^T (\underline{y} - X\underline{\hat{\beta}}) \end{aligned}$$

One way to proceed is to take derivatives ($\frac{\partial}{\partial \underline{x}} A\underline{x} = \underline{x}$, and $\frac{\partial}{\partial \underline{x}} \underline{x}^T A \underline{x} = 2A\underline{x}$) and equate to zero:

$$\begin{aligned} & \frac{\partial}{\partial \underline{\hat{\beta}}} \left[(\underline{y} - X\underline{\hat{\beta}})^T (\underline{y} - X\underline{\hat{\beta}}) \right] \\ &= \frac{\partial}{\partial \underline{\hat{\beta}}} (\underline{y} - X\underline{\hat{\beta}})^T \cdot 2(\underline{y} - X\underline{\hat{\beta}}) \\ &= -2X^T (\underline{y} - X\underline{\hat{\beta}}) \end{aligned}$$

Setting this last expression equal to zero produces the *normal equations*,

$$X^T(\underline{y} - X\underline{\hat{\beta}}) = X^T E = \underline{0}$$

because this asserts that the columns of X are orthogonal to, or in older mathematical language "normal to", the residuals $E = \underline{y} - X\underline{\hat{\beta}}$. Distributing X^T across the $(\underline{y} - X\underline{\hat{\beta}})$ term we obtain

$$X^T \underline{y} = X^T X \underline{\hat{\beta}}$$

and—assuming that $X^T X$ is an invertible matrix—we can premultiply this equation by $(X^T X)^{-1}$ to obtain

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{y}$$

Note that it is just as easy to obtain the fitted values $\underline{\hat{y}}$:

$$\underline{\hat{y}} = X\underline{\hat{\beta}} = X(X^T X)^{-1} X^T \underline{y}$$

The matrix $H = X(X^T X)^{-1} X^T$ is often called the *hat matrix* for linear regression, because the effect of multiplying y by H is to "put the hat on":

$$\underline{\hat{y}} = Hy$$

Remarks:

- The fact that the fitted values $\underline{\hat{y}}$ are a *linear transformation of the original responses* $\underline{y}$ is *really* what makes linear regression "linear". Nonparametric regression methods such as
  - Kernel smoothing

– Spline smoothing
– *but not loess*

can be shown to be linear smoothers in exactly this sense. Linearity makes obtaining distribution theory, large sample theory, etc., particularly easy.

- Note that
$$\frac{\partial}{\partial \underline{y}}\underline{\hat{y}} = H$$
and in particular
$$\frac{\partial}{\partial y_i}\hat{y}_i = h_{ii},$$
the i^{th} diagonal element of the hat matrix H. Thus, the larger h_{ii} is, the more sensitive $\hat{y}_i$ is to changes in y_i; this further explains why h_{ii} is a good place to start for influence diagnostics.

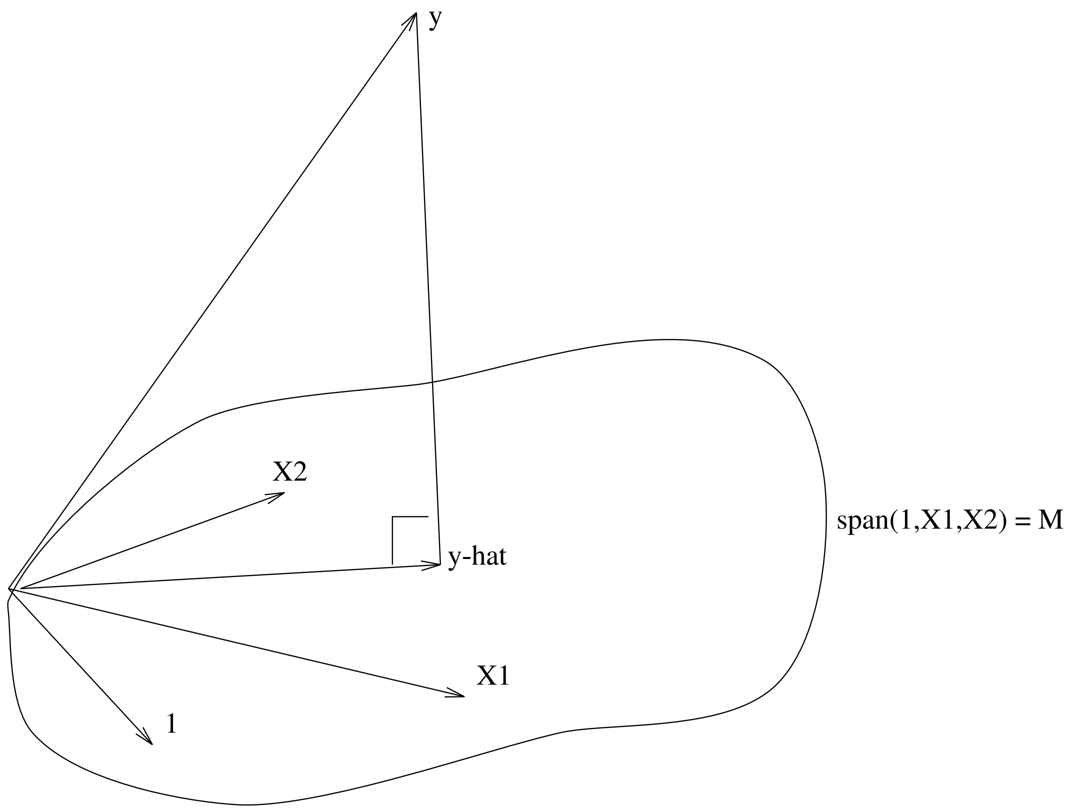
- We can also obtain the residuals using H:
$$\underline{E} = \underline{y} - \underline{\hat{y}} = \underline{y} - H\underline{y} = (I - H)\underline{y}$$

- H has many remarkable properties. Let
$$\begin{aligned} P &= H \\ Q &= 1 - H \end{aligned}$$
Then it is not hard to show, using the definition of H and/or the normal equations, that
$$\begin{array}{rrl} P^T = P, & Q^T = Q: & \text{symmetric} \\ P^2 = P, & Q^2 = Q: & \text{idempotent} \\ PQ = 0: & & \text{orthogonal} \\ P + Q = I: & & \text{complete} \end{array}$$
and indeed P and Q are "projection matrices" in n-dimensional real space: every vector x in $\Re^n$ can be written uniquely as an orthogonal decomposition
$$x = x_P + x_Q$$
where $x_P = Px \in \text{ColSpace}\,(P)$, $x_Q = Qx \in \text{ColSpace}\,(Q)$, and $x_P \perp x_Q$ (they are orthogonal).

- *Calculating the projection and properties of the projection matrix are important, but the algebraic formula $H = X(X^TX)^{-1}X^T$ involving matrix inversion need not be the best computational formula. See Thisted's Statistical Computing text for linear regression estimation formulae that are actually useful computationally.*

2 Some Geometry

Consider a dependent variable $\underline{y} = (y_1, \ldots, y_n)^T$ and $k = 2$ independent variables,
$$\underline{X_1} = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \quad \underline{X_2} = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix}$$



We wish to consider the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$. Letting

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = (\underline{1}, \underline{X_1}, \underline{X_2}),$$

$\underline{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, and $\underline{\beta} = (\beta_1, \beta_2, \beta_3)^T$, we may again write this in matrix form as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

The least-squares fitted values $\underline{\hat{y}}$ minimize the squared distance $\|\underline{y} - \underline{\hat{y}}\|^2$, and so, looking at the picture and thinking about elementary geometry, it's clear that the vector $\underline{E} = \underline{y} - \underline{\hat{y}}$ should be perpendicuar to $\mathcal{M} = \text{Span}\,(\underline{1}, \underline{X_1}, \underline{X_2}) = \text{ColSpace}\,(X)$, the vector space spanned by the columns of X.

On the other hand, if $X^T X$ is of full rank, then the algebra we did above asserts that the least squares estimates satisfy

$$\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{y}$$

and the fitted or predicted values for the y's satisfy

$$\underline{\hat{y}} = Hy = X(X^T X)^{-1} X^T \underline{y} = X\underline{\hat{\beta}}$$

This equation shows immediately that $\underline{\hat{y}} \in \text{ColSpace}\,(X)$, and since $\underline{E} = (I - H)\underline{y}$ we also know that, for any $\underline{\beta}$ (least-squares estimates or not)

$$(X\underline{\beta})^T \underline{E} = \underline{\beta}^T X^T (I - H)\underline{y} = \underline{\beta}^T (X^T - X^T X (X^T X)^{-1} X^T) = \underline{\beta}^T (X^T - X^T) = \underline{0}$$

so that the residuals $\underline{E} = \underline{y} - \underline{\hat{y}}$ really are orthogonal to every possible linear combination of the columns of X.

3 Assumptions and Standard Error Calculations

All of the preceding was just algebra; lets do some statistics. The model is

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

and we assume that

1. X is of full rank (this is easily fixed)
2. $E(\underline{\epsilon}) = \underline{0}$
3. $\text{Var}(\underline{\epsilon}) = \sigma^2 \underline{I}_{n \times n}$ which incorporates the equal variance and zero covariance assumptions
4. $\underline{\epsilon} \sim N(\underline{0}, \sigma^2 \underline{I})$ a multivariate normal distribution.

3.1 Digression — Multivariate Normal Distribution

Definition

A random vector

$$\underline{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

is said to have a multivariate normal distribution $N(\underline{\mu}, \Sigma)$ if

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}, \text{ and}$$

$$E(\underline{x}) = \underline{\mu}, \text{ and}$$

$$\text{Cov}(x_i, x_j) = \sigma_{i,j}$$

AND the distribution of

$$\underline{a}^T \underline{x} \;\text{ is }\; N(\underline{a}^T \underline{\mu}, \underline{a}^T \Sigma \underline{a})$$

for *all* vectors $\underline{a}$.

In other words *it is multivariate normal iff every possible linear combination is univariate normal.*

Transformation Properties

If $\underline{x} \sim N_n(\underline{\mu}, \Sigma)$ and $W_{p\times n}$ is a matrix of constants then

$$W\underline{x} \sim N_p(W\underline{\mu}, W\Sigma W^T)$$

Now let $\Sigma_{n\times n}$ be decomposed as $V_{n\times n}V_{n\times n}^T$, where each V is of full rank (called *a* square root of W. This can always be done (in many ways). Then

$$V^{-1}\underline{x} \sim N_n(V^{-1}\underline{\mu},\ \underbrace{V^{-1}(VV^T)V^{-T}}_{I_{n\times n}})$$

Thus, if $\underline{x}$ is a random vector with mean μ and variance Σ and $\Sigma = VV^T$, then $V^{-1}\underline{x}$ has mean $V^{-1}\underline{\mu}$ and variance I.

3.2 Sampling Distributions of $\underline{\hat{\beta}}$ and $\underline{\hat{y}}$

We are now in the position to say something about $\underline{\hat{\beta}} = (X^TX)^{-1}X^T\underline{y}$ and $\underline{\hat{y}} = H\underline{y}$.

For starters, note that

$$\begin{aligned} E(\underline{y}) &= X\underline{\beta}, \text{ and} \\ \text{Var}(\underline{y}) &= \sigma^2 \cdot I. \end{aligned}$$

Thus

$$\begin{aligned} E(\underline{\hat{\beta}}) &= (X^TX)^{-1}X^TE(\underline{y}) \\ &= (X^TX)^{-1}X^TX\underline{\beta} \quad = \quad \underline{\beta} \end{aligned}$$

and

$$\begin{aligned} \text{Var}(\underline{\hat{\beta}}) &= (X^{-T}X)^{-1}X^T\text{Var}(\underline{y})X(X^{-T}X)^{-1} \\ &= (X^{-T}X)^{-1}X^T(\sigma^2 \cdot I)X(X^{-T}X)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned}$$

The fitted values $\underline{\hat{y}}$ and residuals $\underline{E}$ are even easier

$$\begin{aligned} E[\underline{\hat{y}}] &= E[H\underline{y}] \ = \ HX\underline{\beta} \ = \ X\underline{\beta} \ = \ E[\underline{y}] \\ E[\underline{\hat{E}}] &= E[(I-H)\underline{y}] \ = \ E[\underline{y}] - E[H\underline{y}] \ = \ E[\underline{y}] - E[\underline{y}] \ = \ \underline{0} \end{aligned}$$

and

$$\begin{aligned} \text{Var}\,(\underline{\hat{y}}) &= \text{Var}\,(H\underline{y}) \ = \ H\text{Var}\,(\underline{y})H^T \ = \ H\sigma^2IH^T \ = \ \sigma^2HH^T \ = \ \sigma^2H \\ \text{Var}\,(\underline{E}) &= \text{Var}\,((I-H)\underline{y}) \ = \ \cdots \ = \ \sigma^2(I-H) \end{aligned}$$

Together with the assumption that $\underline{\epsilon} \sim N(\underline{0}, \sigma^2I)$, or $\underline{y} \sim N(X\underline{\beta}, \sigma^2I)$, these facts show that

$$\boxed{\underline{\hat{\beta}} \sim N(\underline{\beta}, \sigma^2(X^TX)^{-1}) \qquad \underline{\hat{y}} \sim N(X\underline{\beta}, \sigma^2H) \qquad \underline{E} \sim N(\underline{0}, \sigma^2(I-H))}$$

How do we estimate σ^2? From our earlier work we expect that we can get an unbiased estimate as a multiple of RSS. To calculate $E[RSS]$ we need a few preliminaries.

First we need a fact from linear algebra, namely that if both AB and BA are legal matrix multiplications, then

$$\text{trace}\,(AB) = \text{trace}\,(BA)$$

where $\text{trace}\,(A) = \sum_i a_{ii}$ is the sum of the diagonal elements of A. We immediately put this to good use, to see that

$$\text{trace}\,(H) = \text{trace}\,(X(X^TX)^{-1}X^T) = \text{trace}\,(X^TX(X^TX)^{-1}) = \text{trace}\,(I_{(k+1)\times(k+1)}) = k+1$$

Next we need a result about the expected value of a quadratic form

$$E[\underline{y}^T A\underline{y}] = \text{trace}\,(A\text{Var}\,(\underline{y})) + E[\underline{y}]^T AE[\underline{y}]$$

(this is analogous to the univariate fact that $E[cy^2] = c^2\sigma^2 + c^2E[y]^2$). Finally we note that

$$\begin{aligned} RSS &= (\underline{y} - \underline{\hat{y}})^T(\underline{y} - \underline{\hat{y}}) \\ &= \underline{y}^T[I-H]^T[I-H]\underline{y} \\ &= \underline{y}^T[I-H]\underline{y} \end{aligned}$$

With these three bits of information we calculate

$$\begin{aligned} E(RSS) &= E[\underline{y}^T[I-H]\underline{y}] \\ &= \text{trace}\,([I-H]\sigma^2 I) + (X\beta)^T(I-H)(X\beta) \\ &= \sigma^2\text{trace}\,(I-H) + \underline{0} \\ &= \sigma^2[n-(k+1)] \end{aligned}$$

Consequently,

$$E\left[\frac{RSS}{n-(k+1)}\right] = \sigma^2$$

so that

$$\hat{\sigma}^2 = \frac{RSS}{n-(k+1)}$$

is an unbiased estimate of σ^2, as we expected.

3.3 Variances of Fitted Values and Predicted Values

We have done part of this already. Note that

$$\begin{aligned} \hat{y}_i &= \underline{X_i}^T \cdot \underline{\hat{\beta}} \\ &= \underline{X_i}^T(X^TX)^{-1}X^T\underline{y} \end{aligned}$$

so that

$$\text{Var}(\hat{y}_i) = \sigma^2\underline{X_i}^T(X^TX)^{-1}\underline{X_i}$$

Notice that the variance of the fitted value depends on the vector X where y_i was observed.

At a new value $\underline{x}^*$ that was not in the original data set, the fitted value is still

$$\hat{y}^* = \underline{x^*}^T \underline{\hat{\beta}}$$

with variance

$$\text{Var}\,(\hat{y}^*) = \sigma^2 \underline{x^*}^T (X^T X)^{-1} \underline{x^*}$$

Note that $\underline{x^*}$ does not have to be part of the original X matrix. The variance calculated here is for estimating $E[y^*]$ in the equation

$$y^* = \underline{x^*}^T \beta + \epsilon^*$$

Sometimes it is desirable to try to predict the actual value y^* itself, and not just the expected value $E[y^*]$. Assuming as usual that ϵ^* is independent of the other ϵ's in the problem, and comes from the same $N(0, \sigma^2)$ distribution, we have simply

$$\begin{aligned}
\text{Var}\,(\hat{y}^*_{pred}) &= \text{Var}\,(\hat{y}^*) + \text{Var}\,(\epsilon^*) \\
&= \sigma^2 \underline{x^*}^T (X^T X)^{-1} \underline{x^*} + \sigma^2 \\
&= \sigma^2 \left[1 + \underline{x^*}^T (X^T X)^{-1} \underline{x^*}\right]
\end{aligned}$$

4 Some final remarks

Aside from being a clean way to derive and generalize all of the standard formulas from simple regression, there are several reasons for bothering with the linear algebra formulation:

- The geometry suggested above is sometimes quite helpful in figuring out what vague notions in regression mean. For example, "collinearity" of $\underline{X_1}$ and $\underline{X_2}$ simply means that the angle between $\underline{X_1}$ and $\underline{X_2}$ is so small that the position of the hyperplane Span $(\underline{1}, \underline{X_1}, \underline{X_2})$ is very sensitive to small changes in $\underline{X_1}$ or $\underline{X_2}$,
- Standard error calculations for the β's are clearly straightforward (though messy calculations are hidden in the matrix multiplications!). Since we also have expressions for the covariances, we can easily obtain standard errors for any linear combination (contrast!) of the parameters.
- The geometry is also useful for suggesting what is the "same" model—any parametrization that generates the same H matrix is the "same" model as far as linear regression is concerned.
- Finally the geometry, mixed with a little distribution theory, will be very helpful in creating and understanding standard errors that are corrected for multiple inference, such as in Scheffé's method.

You may also be interested in knowing that, just as in the simple regression case, the least-squares estimates are the same as the maximum likelihood estimates of the β's; and as before the MLE of σ^2 is the biased estimate RSS/n.