## MULTIVARIATE LINEAR REGRESSION MODEL

Consider a data set where we have *n* observations (cases) on *m* outcomes (dependent variables) with *p* explanatory variables (independent variables).

This is where we have multiple dependent variables, so the dependent variables are represented by a matrix, **Y**, which has *n* (cases) rows and *m* (variables) columns. **Y** is an n × m matrix.

Because of this, our regression coefficients, betas $\boldsymbol{\beta}$, are also in matrix form, since now there are beta coefficients for each predictor and for each outcome variable in the model. Similarly, we have a matrix of residuals, $\boldsymbol{\varepsilon}$, since there are residuals for each person on each outcome variable.

$$\underset{n \times m}{\mathbf{Y}} = \underset{n \times (p+1)}{\mathbf{X}} \quad \underset{(p+1) \times m}{\boldsymbol{\beta}} \quad + \quad \underset{n \times m}{\boldsymbol{\varepsilon}}$$

***Multiple Outcomes***

In most social science and educational settings, there are naturally multiple outcome variables (dependent variables). We typically do not engage in research to study a single outcome, although this might be the case. However, it’s more typical to study multiple outcomes.

When we study academic achievement, we often look at multiple outcomes of achievement, including reading, writing, mathematics, and science. Or, we might look at multiple components of mathematics, including algebra, probability, and geometry. In social science settings, we may look at multiple components of personality, including neuroticism, extroversion, and openness – or even the big five personality factors. When we are interested in examining motivation, we may include activation, persistence, and intensity.

The issue in empirical research, with a general linear model approach, is that these outcomes are correlated. If we examine each outcome with an independent model, results are not necessarily independent, which increases the study-wise Type-I error rate; we would suggest more statistically significant effects than are likely to exist in the population because of the dependence across the independent GLMs.

Multivariate statistics include a class of models that simultaneously include multiple variables, typically considering those variables as outcomes or dependent variables. This is different than a model like multiple regression – where there are multiple predictors. Multivariate regression includes the case where there are multiple outcomes, typically as well as multiple predictors.

Multivariate statistical models include multivariate regression and multivariate analysis of variance (MANOVA) – otherwise known as GLM. Other common models include factor analysis, principal components analysis, discriminant analysis, canonical correlations, cluster analysis, and others.

The matrix results we observed for regression and the general linear model with a single outcome dependent variable work in the multivariate context. Given the multivariate GLM:

$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

We can solve for the regression coefficient matrix:

$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

$\varepsilon = \mathbf{Y} - \mathbf{X}\beta$

$\varepsilon'\varepsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$ ← error SSCP

***OLS Solution***

In the typical GLM, our task is to find the ordinary least-squares solution. This means that our task is to minimize the sum of the squared residuals. The OLS estimators are best linear unbiased estimates (BLUE) when the errors are homoscedastic and serially uncorrelated.

The OLS task in the multivariate model is to minimize the trace of the error SSCP matrix, $\varepsilon'\varepsilon$:

$$\text{trace}[(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)]$$

Also, OLS estimates of $\beta$ minimize the generalized variance (determinant):

$$|(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)|$$

We can then consider the decomposition of variance or sums of squares:

There is a predicted value for each outcome. The predicted values vector is:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

There is a residual for each outcome. The residual vector is:

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I}\text{-}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{Y}$$

The sums of squares are:

$\mathbf{Y}'\mathbf{Y} = (\hat{\mathbf{Y}} + \hat{\varepsilon})'(\hat{\mathbf{Y}} + \hat{\varepsilon}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} + \hat{\varepsilon}'\hat{\varepsilon} + \mathbf{0} + \mathbf{0} = \text{SSR} + \text{SSE} = \text{SST}$