Stats refresher

Variance

Expectation:

$$\sigma_x^2 = E[(x - \mu_x)^2]$$

Sampling estimator for variance of x:

$$\hat{\sigma}_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Covariance

Expectation:

$$\sigma_{x,y}^2 = E[(x - \mu_x)(y - \mu_y)]$$

Sampling estimator for covariance of x and y:

$$\hat{\sigma}_{x,y} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

The covariance of a variable with itself is the variance of that variable.

Correlation

Correlation provides a dimensionless metric for the covariance. It is the ratio of covariance in x and y relative to the square-root of the product of their variances.

$$r_{x,y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Ordinary least squares derivation

The goal of ordinary least squares is to minimize the sum of the squared residuals, or is to **maximize the variation in y that is explained by x**.

The cost function for OLS is given by

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{N} e_i^2$$

$$\min_{\beta_0,\beta_1}\sum_{i=1}^{N}(y_i - \beta_0 + \beta_1 x_i)^2$$

We can find the minima of this cost function by taking the partial derivatives with respect to β_0 and β_1 and setting them to zero. For β_0 the partial derivative is

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial \beta_0} = \sum_{i=1}^{N} -2(y_i - \beta_0 - \beta_1 x_i) = 0$$

and for β_1

$$\frac{\partial \sum_{i=1}^{N} e_i^2}{\partial \beta_0} = \sum_{i=1}^{N} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

These can be solved pretty easily. We know

$$\begin{gathered}
\sum_{i=1}^{N} \beta_1 x_i = N\beta_1 \bar{x} \\
\sum_{i=1}^{N} \beta_1 y_i = N\bar{y} \\
\sum_{i=1}^{N} \beta_1 x_i = N\bar{x} \\
\sum_{i=1}^{N} \beta_0 x_i = N\beta_0
\end{gathered}$$

Starting with β_0, we can remove the -2 and substitute the equalities above giving

$$N\bar{y} - N\beta_0 - N\beta_1 \bar{x} = 0$$

Rearranging this and solving for β_0 gives

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

From this we can see that **the regression line will pass through the means of x and y.**

Similar to β_0, for β_1 we drop the -2, rearrange the equation and substitute the solution for β_0

$$
\begin{aligned}
\sum_{i=1}^{N} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \\
\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \beta_0 - \sum_{i=1}^{N} \beta_1 x_i^2 = 0 \\
\sum_{i=1}^{N} x_i y_i - (\bar{y} - \beta_1 \bar{x}) x_i - \beta_1 x_i^2 = 0 \\
\sum_{i=1}^{N} y_i x_i - \bar{y} \sum_{i=1}^{N} x_i + \beta_1 \bar{x} \sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \\
\sum_{i=1}^{N} x_i y_i - N\bar{y}\bar{x} + N\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{N} x_i^2 = 0
\end{aligned}
$$

We can substitute and solve for β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}$$

If we know

$$
\begin{aligned}
\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - N\bar{x}^2 \\
\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}
\end{aligned}
$$

Then the solution for β_1 becomes

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

Thus, the slope is the the covariance of x and y relative to the variance of x.

In addition to the bold-faced text above, **the predictor and residuals are uncorrelated.** The covariance between a constant some variable will always be 0. The covariance of a sum is equal to the sum of the covariances of the elements.

$$\begin{aligned}
Cov(x,e) &= Cov[x,(y-\beta_0+\beta_1 x)] \\
&= Cov(x,y) - Cov(x,\beta_0) + \beta_1 Cov(x,x) \\
&= Cov(x,y) - 0 + \frac{Cov(x,y)}{Var(x)} Cov(x,x) \\
&= Cov(x,y) - Cov(x,y) = 0
\end{aligned}$$

Matrix algebra

We denote a matrix using bold capital letters and vectors using bold lowercase letters. The dimensions of a matrix are presented as the number of rows by the number of columns. The cells of a matrix or vector are referred to as elements. If we pull a column or row out of a matrix it is a vector.

Matrix addition and subtraction

Addition of two matrices requires that the matrices have the same dimensions. Addition is done element-wise. This is the same for subtraction.

$$\mathbf{C} = \mathbf{A} + \mathbf{B}$$
$$\begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 4 \\ 4 & 7 \end{bmatrix}$$

Matrix multiplication

A scalar multiplied by a matrix is simply the product of each element of the matrix by the scalar.

$$\mathbf{C} = a\mathbf{B}$$
$$\begin{bmatrix} 2 & 8 \\ 8 & 14 \end{bmatrix} = 2 \begin{bmatrix} 1 & 4 \\ 4 & 7 \end{bmatrix}$$

Multiplying two matrices or vectors is a bit more involved. The product of two vectors is the sum of the products of each of their elements (the dot product). For instance,

$$\mathbf{ab} = \sum_{i=1}^{n} a_i \cdot b_i$$

This requires that both vectors have the same number of elements.

The same can be done applied to the multiplication of two matrices (**AB**). Each matrix is split into vectors (first matrix is split into row vectors and the second to column vectors). Each element of the resulting matrix is then calculated via

$$\mathbf{C} = \mathbf{AB}$$
$$C_{i,j} = \mathbf{a}_i \cdot \mathbf{b}_j = \sum_{k=1}^{c} a_{ik} \cdot b_{jk}$$

i refers to a row, j to a column and c is the number of columns in $\mathbf{A}$ or the number of rows in $\mathbf{B}$. This is only possible if the number of columns in $\mathbf{A}$ is equal to the number of rows in $\mathbf{B}$.

Transpose of a matrix

To transpose a matrix we simply swap the rows and columns. Transpose is denoted using ^{T} or $'$.

$$\mathbf{C} = \begin{bmatrix} 1 & 2 \\ 4 & 7 \end{bmatrix}$$
$$\mathbf{C}^T = \begin{bmatrix} 1 & 4 \\ 2 & 7 \end{bmatrix}$$

Inverse of a matrix

This is analogous to division of scalars. Before we get into computing inverses we need to have an understanding of determinates and minors.

Determinates

Determinate are scalars calculated from a square matrix, and is denoted by $|\mathbf{A}|$ or $\det(\mathbf{A})$ for the matrix $\mathbf{A}$. This is important because the inverse of a square matrix exists *if and only if* $|\mathbf{A}| \neq 0$.
For a 1×1 matrix,

$$|\mathbf{X}| = \mathbf{X}$$

For a 2×2 matrix,

$$\mathbf{X} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$
$$|\mathbf{X}| = ad - cb$$

For a 3×3 matrix,

$$\mathbf{X} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$
$$|\mathbf{X}| = a(ei - fh) - b(di - fg) + c(dh - ge)$$

This can be summarized as

$$|\mathbf{X}| = \sum_{j=2}^{n} X_{ij}(-1)^{i+j}|\mathbf{X}_{ij}|$$

$\mathbf{X}_{ij}$ is a minor of $\mathbf{X}$ with the i^{th} row and j^{th} removed and X_{ij} is an element of the matrix $\mathbf{X}$.

Calculating the inverse of a matrix

The inverse of a matrix $\mathbf{X}$ is calculated as

$$X_{ij}^{-1} = \left[\frac{(-1)^{i+j}|\mathbf{X}_{ij}|}{|\mathbf{X}|}\right]^T$$

The inverse acts like scalar division. For instance, just as $aa^{-1} = 1$, the product of a square matrix and its inverse is a square matrix with 1's along the diagonal and 0's in the off-diagonal. This resulting matrix is called an identity matrix and is denoted as $\mathbf{I}$.

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Why is this important?

Matrix algebra allows us to express a system of equations and their solutions in a compact form. This will become apparent below as we express the OLS derivation in matrix form.

OLS in matrix form and its derivation

From above, we can express each observation in our experiment as the sum of the mean, some linear combination of predictors and some random noise.

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

This can also be expressed in a more compact matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

. The coefficients are now stored in the vector $\boldsymbol{\beta}$ and the matrix $\mathbf{X}$ contains the predictor as well as a vector of 1's for the intercept.

For simplicity, let's assume that we have centered the response vector $\mathbf{y}$ and the predictor. Now our regression will pass though the origin and the intercept $\beta_0 = 0$. The OLS solutions for the coefficients are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

Since $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{N} x_i^2$, we can express the denominator as

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

And the numerator can be expressed as

$$\mathbf{X}'\mathbf{y}$$

Thus, the OLS solution in matrix form is

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

And the variance of these estimates is

$$(\mathbf{X}'\mathbf{X})^{-1}\sigma_e^2$$