Deriving Classic Results in Linear Regression with The Algebra of Variances and Covariances

Earlier in the course, we derived some key results regarding the variances and covariances of linear transformations and linear combinations. In this handout, we revisit these results and re-express them as a succinct set of rules for manipulating and deriving variances and covariances.

Recall that early in the course, we first pointed out that if Y is a linear transformation of X, i.e., $Y = aX + b$, then $S_Y^2 = a^2 S_X^2$. There's a bit more to this result than meets the eye at first, because we also proved that if $Y = aX + b$, $dy = adx$, that is, additive constants never affect deviation scores, so of course they cannot affect variances or covariance, which are functions of deviation scores.

Consequently, we can state our first rules of covariance algebra. Let $S_{A,B}$ stand for the covariance between variables A and B. We recall that the variance of any variable is its covariance with itself, so

$$S_A^2 = S_{A,A} \tag{1}$$

We developed a general heuristic rule for deriving variances and covariances of linear transformations and linear combinations. That rule, in a nutshell, says

1. Write the expression(s) whose variance(covariance) you wish to derive.
2. Take the square(cross-product) of the expression(s).
3. Apply the following conversion rules:
   (a) Constants that are multiplied by variables are left unchanged, but lone additive or subtractive constants may be removed prior to processing the expression
   (b) Any squared variable is converted to the variance of that variable.
   (c) Any product of two variables is converted to the covariance of those variables.
   (d) Any expression not containing either the square of a variable or product of two variables is dropped.

These 4 examples demonstrate the essential characteristics of the heuristic rules:

1. *Derive the variance of* $X + Y$. We square $X + Y$, getting $X^2 + Y^2 + 2XY$, Applying the conversion rule, we get $S_X^2 + S_Y^2 + 2S_{X,Y}$.
2. *Derive the covariance of* $X + Y$ *and* $X - Y$. Taking the cross-product, we get $(X + Y)(X - Y) = X^2 - Y^2$. Applying the conversion rule, we get $S_X^2 - S_Y^2$.

3. *Derive the variance of* $aX + b$. We can proceed in two distinct ways. In one approach, we drop the b first, since it is a lone additive constant. We then square aX, obtaining a^2X^2, and then we apply the conversion rule, getting $a^2S_X^2$. Alternatively, we could square the original expression, getting $(aX + b)^2 = a^2X^2 + b^2 + 2abX$. Applying the conversion rule, we drop the last two terms, which do not have either the square of a variable or the product of two variables, and we end up with $a^2S_X^2$ again.

4. *Derive the covariance of* $aX + b$ *and* cY. Taking the product of the two expressions, we get $(aX + b)(cY) = acXY + bcY$. Applying the conversion rule, we drop the second term, because it does not have the square of a variable or the product of two variables. The result is thus $acS_{X,Y}$.

The results of the heuristic rule may be expressed very succinctly with the following notational variation. For variables X and Y and constants a, b, c, d, we have the following:

1. $$S_{aX+b,cY+d} = S_{aX,bY} = abS_{X,Y} \tag{2}$$

2. $$S^2_{aX+b} = S^2_{aX} = a^2S^2_X \tag{3}$$

3. $$S^2_{aX+bY} = a^2S^2_X + b^2S^2_Y + 2abS_{X,Y} \tag{4}$$

4. $$S_{aX+bY,cX+dY} = acS^2_X + bdS^2_Y + (bc + ad)S_{X,Y} \tag{5}$$

The above notation and the results it describes can be used to produce very succinct derivations of some classic results in linear regression.

Consider the simple least-squares linear regression setup for predicting Y from X. The key equations are

$$\begin{aligned} Y_i &= \beta_0 + \beta_1X_i + E_i \qquad (6) \\ &= \hat{Y}_i + E_i \qquad (7) \end{aligned}$$

where, of course

$$E_i = Y_i - \hat{Y}_i \tag{8}$$

and $\hat{Y}_i$ is defined as

$$\hat{Y}_i = \beta_1X_i + \beta_0 \tag{9}$$

The values of β_1 and β_0 that minimize the sum of squared residuals are

$$\beta_1 = r_{Y,X}\frac{S_Y}{S_X} = \frac{S_{Y,X}}{S^2_X} \tag{10}$$

and

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \tag{11}$$

We'll now prove the results given in the lecture notes.

First, *prove that the variance of the predicted scores is given by*

$$S^2_{\hat{Y}} = r^2_{X,Y} S^2_Y \tag{12}$$

Here is the succinct proof. First, via substitution, we write

$$S^2_{\hat{Y}} = S^2_{\beta_1 X + \beta_0} \tag{13}$$

We then apply the above algebraic results, beginning by dropping the constant β_0.

$$S^2_{\beta_1 X + \beta_0} = S^2_{\beta_1 X} = \beta_1^2 S^2_X \tag{14}$$

Next, we simply substitute Equation 10, obtaining

$$\beta_1^2 S^2_X = \left(r_{Y,X} \frac{S_Y}{S_X} \right)^2 S^2_X = r^2_{Y,X} \frac{S^2_Y}{S^2_X} S^2_X = r^2_{Y,X} \frac{S^2_Y}{\cancel{S^2_X}} \cancel{S^2_X} = r^2_{Y,X} S^2_Y \tag{15}$$

The second result we seek to prove is that *predicted and error scores are always uncorrelated*, i.e.,

$$S_{\hat{Y},\ E} = 0 \tag{16}$$

This result falls out directly from substitution. The key is to choose the more opportune version of the formula for β_1. We start with a direct substitution:

$$S_{\hat{Y},\ E} = S_{\beta_1 X + \beta_0,\ Y - \beta_1 X - \beta_0} \tag{17}$$

Next we drop the additive constants

$$S_{\beta_1 X + \beta_0,\ Y - \beta_1 X - \beta_0} = S_{\beta_1 X,\ Y - \beta_1 X} \tag{18}$$

From here, the manipulations involve straightforward (if slightly messy) substitution of definitions:

$$\begin{aligned}
S_{\beta_1 X,\ Y - \beta_1 X} &= \beta_1 S_{X,Y} - \beta_1^2 S^2_X && (19)\\
&= \beta_1 S_{Y,X} - \beta_1^2 S^2_X && (20)\\
&= \frac{S_{Y,X}}{S^2_X} S_{Y,X} - \left(\frac{S_{Y,X}}{S^2_X} \right)^2 S^2_X && (21)\\
&= \frac{S_{Y,X}}{S^2_X} S_{Y,X} - \frac{S^2_{Y,X}}{S^4_X} S^2_X && (22)\\
&= \frac{S^2_{Y,X}}{S^2_X} - \frac{S^2_{Y,X}}{S^2_X} && (23)\\
&= 0 && (24)
\end{aligned}$$

Finally, we show that $S_E^2 = (1 - r_{X,Y}^2)S_Y^2$. Note that we have just shown that $\hat{Y}$ and E have a zero covariance. Since $Y = \hat{Y} + E$, this immediately implies

$$S_Y^2 = S_{\hat{Y}+E}^2 = S_{\hat{Y}}^2 + S_E^2 + \cancel{2S_{\hat{Y},E}} = S_{\hat{Y}}^2 + S_E^2 \tag{25}$$

Using the previous result for $S_{\hat{Y}}^2$, we get

$$S_Y^2 = S_{\hat{Y}}^2 + S_E^2 = r_{X,Y}^2 S_Y^2 + S_E^2 \tag{26}$$

By subtraction, this implies that

$$S_E^2 = S_Y^2 - r_{X,Y}^2 S_Y^2 = (1 - r_{X,Y}^2)S_Y^2 \tag{27}$$

We have, therefore, completed the proof of the three results in the lecture notes.

In class, we then went on to show that these results imply that $r_{X,Y}^2$ is the proportion of total squared error that is eliminated by adding the term $\beta_1 X$ to the regression equation $\hat{Y}_i = \beta_0$.

Specifically, suppose we did not know X and had to predict the Y scores with the equation $\hat{Y} = \beta_0$. The best β_0, the one that minimizes the sum of squared errors, can be shown (we proved the result in class) to be $\bar{Y}$, the sample mean of the Y scores. In that case, the sum of squared errors is simply the sum of squared deviations of the Y scores around their mean, which may be written

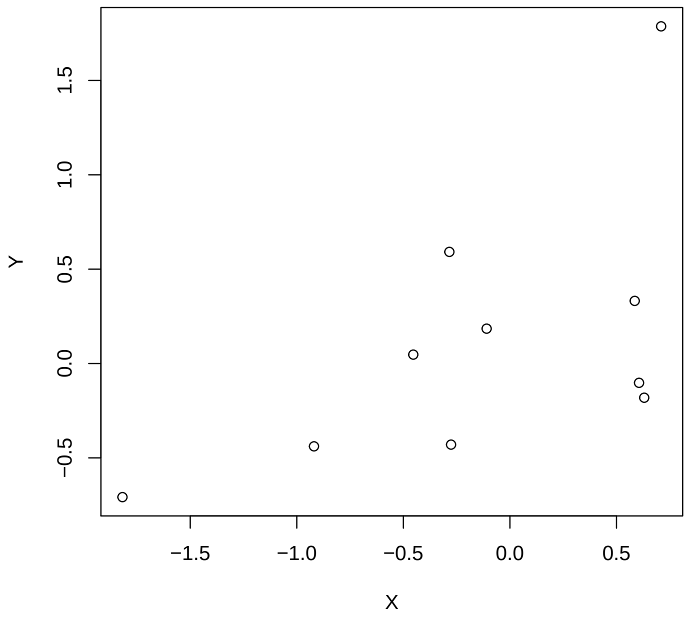
$$\sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n}(Y_i - \beta_0)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (n-1)S_Y^2 \tag{28}$$

On the other hand, the sum of squared errors when the full regression equation is used is $(n-1)S_E^2 = (n-1)(1 - r_{X,Y}^2)S_Y^2$. The amount saved is the difference between these two values, i.e., $(n-1)S_Y^2 - (n-1)(1 - r_{X,Y}^2)S_Y^2 = (n-1)r_{Y,X}^2 S_Y^2$. The *proportion* of the maximum squared error saved by using X is thus

$$\frac{(n-1)r_{X,Y}^2 S_Y^2}{(n-1)S_Y^2} = r_{X,Y}^2 \tag{29}$$

To illustrate the above results, let's analyze a small artificial data set. We create two variables, X and

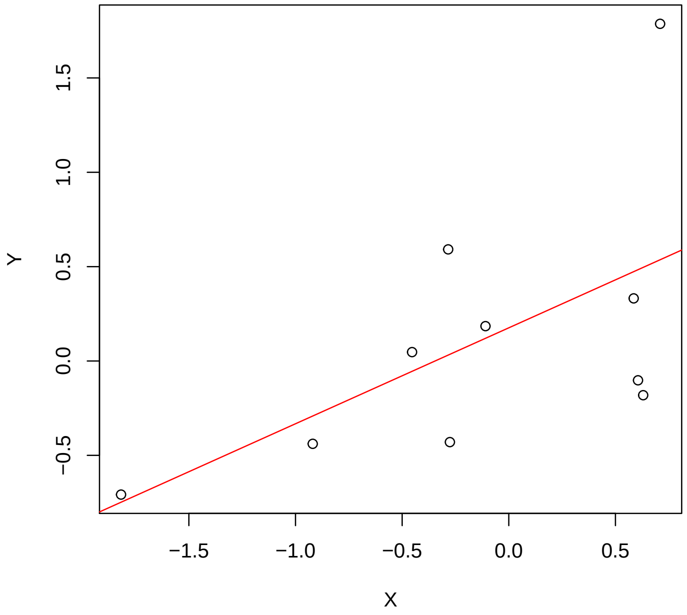
```
> set.seed(12345)
> X <- rnorm(10)
> Y <- sqrt(1/2)*X + sqrt(1/2)*rnorm(10)
> plot(X,Y)
```



Fitting the simple linear regression model, we get

```
> fit.1 <- lm(Y ~ X)
> plot(X,Y)
> summary(fit.1)

Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67730 -0.38568 -0.05055  0.09258  1.25025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1758     0.1950   0.902   0.3936
X             0.5083     0.2489   2.042   0.0754 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6076 on 8 degrees of freedom
Multiple R-squared: 0.3427,       Adjusted R-squared: 0.2605
F-statistic: 4.171 on 1 and 8 DF,  p-value: 0.07541
```

```
> abline(fit.1,col="red")
```



The predicted and residual scores can be obtained directly from the fit object by use of the `predict()` and `residuals` functions. Below, we calculate the predicted and error scores, then show that their variances and covariance match the theory we just derived. The covariance between $\hat{Y}$ and E is within rounding error of zero.

```
> Yhat <- predict(fit.1)
> E <- residuals(fit.1)
> var(Yhat)

[1] 0.1710815

> r.YX <- cor(Y,X)
> r.YX^2 * var(Y)

[1] 0.1710815

> var(E)

[1] 0.3281642

> (1 - r.YX^2) * var(Y)
```

```
[1] 0.3281642
```

```
> cov(Yhat,E)
```

```
[1] -4.635604e-17
```

```
> data <- cbind(Y,X,Yhat,E)
> data
```

```
             Y          X        Yhat           E
1   0.33183179  0.5855288  0.47344529 -0.14161350
2   1.78670190  0.7094660  0.53644850  1.25025340
3   0.18478436 -0.1093033  0.12022887  0.06455549
4   0.04717766 -0.4534972 -0.05474133  0.10191899
5  -0.10227913  0.6058875  0.48379456 -0.58607369
6  -0.70785358 -1.8179560 -0.74836101  0.04050743
7  -0.18120246  0.6300986  0.49610222 -0.67730467
8  -0.42975242 -0.2761841  0.03539539 -0.46514780
9   0.59153223 -0.2841597  0.03134099  0.56019125
10 -0.43882927 -0.9193220 -0.29154238 -0.14728689
```

The sum of squared errors when using the full regression equation is

```
> SS.full <- sum(E^2)
> SS.full
```

```
[1] 2.953478
```

Now suppose we fit a model with only an intercept

```
> fit.2 <- lm(Y ~ 1)
> SS.intercept.only <- sum(residuals(fit.2)^2)
> SS.intercept.only
```

```
[1] 4.493211
```

The proportion of squared error saved is calculated below, and exactly matches r^2_{YX}.

```
> (SS.intercept.only - SS.full)/ SS.intercept.only
```

```
[1] 0.34268
```

```
> r.YX^2
```

```
[1] 0.34268
```

As a final comment, I should note that there are notational variations of the algebra of variances and covariances that you will encounter. One obvious variation is to use a $\mathrm{var}(A)$ operator instead of S_A^2, and a $\mathrm{cov}(A, B)$ operator in place of $S_{A,B}$. So, for example, we could write

The results of the heuristic rule may be expressed very succinctly with the following notational variation. For variables X and Y and constants a, b, c, d, we have the following:

1. $$\mathrm{cov}\, aX + b, cY + d = \mathrm{cov}(aX, bY) = ab\,\mathrm{cov}(X, Y) \tag{30}$$

2. $$\mathrm{var}(aX + b) = \mathrm{var}(aX) = a^2\,\mathrm{var}(X) \tag{31}$$

3. $$\mathrm{var}(aX + bY) = a^2\,\mathrm{var}(X) + b^2\,\mathrm{var}(Y) + 2ab\,\mathrm{cov}(X, Y) \tag{32}$$

4. $$\mathrm{cov}(aX+bY, cX+dY) = ac\,\mathrm{var}(X)+bd\,\mathrm{var}(Y)+(bc+ad)\,\mathrm{cov}(X, Y) \tag{33}$$

There is a real advantage to this notational variation. Since the algebra of variances and covariances holds for random variables as well as for lists of numbers, this latter notation can be employed identically for discussions of sample statistics or population results. Another advantage is that it can be directly imported into R.