Lecture 27 – Centered Vectors and the Sample Correlation *R*

Centering

Define the new vectors $\dot{\mathrm{X}} = \mathrm{X} - \bar{\mathrm{X}} = \begin{bmatrix} X_1 - \bar{X} \\ X_2 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}$ and $\dot{\mathrm{Y}} = \mathrm{Y} - \bar{\mathrm{Y}} = \begin{bmatrix} Y_1 - \bar{Y} \\ Y_2 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix}$. The new vectors $\dot{\mathrm{X}}$ and $\dot{\mathrm{Y}}$ are the **Centered** versions of the original vectors X and Y. Centering the data has many advantages and is often done prior to running a regression. I will focus on just a couple of the applications of centering.

Centering and the Sample Correlation

One of the coolest applications of centering involves the sample correlation *r* (or *R*). The sample correlation *R* is an estimator of the correlation between the variables ***X*** and ***Y***, $\rho_{\boldsymbol{X},\boldsymbol{Y}} = \dfrac{Cov(\boldsymbol{X},\boldsymbol{Y})}{\sigma_{\boldsymbol{X}}\sigma_{\boldsymbol{Y}}}$, where $Cov(\boldsymbol{X},\boldsymbol{Y}) = E(\boldsymbol{X} - \mu_X)(\boldsymbol{Y} - \mu_Y)$ is the covariance of ***X*** and ***Y***. The sample correlation *R* is computed as $R = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$. Now, rewriting this using centered vectors, $R = \dfrac{\sum_{i=1}^{n}\dot{X}_i\dot{Y}_i}{\sqrt{\sum_{i=1}^{n}\dot{X}_i{}^2}\sqrt{\sum_{i=1}^{n}\dot{Y}_i^2}}$.

The last equation can be rewritten $R = \dfrac{\dot{\mathrm{X}}^T\dot{\mathrm{Y}}}{\sqrt{\dot{\mathrm{X}}^T\dot{\mathrm{X}}}\sqrt{\dot{\mathrm{Y}}^T\dot{\mathrm{Y}}}} = \dfrac{\dot{\mathrm{X}}^T\dot{\mathrm{Y}}}{\|\dot{\mathrm{X}}\|\|\dot{\mathrm{Y}}\|}$. The final expression may not look familiar unless you've had Multivariate Calculus, so let me introduce the relevant theorem.

Theorem: The dot product of vectors *u* and *v* obeys the equality $u \cdot v = \|u\|\|v\| cos\,\theta$, where θ is the angle between the vectors. The corollary is $cos\,\theta = \dfrac{u \cdot v}{\|u\|\|v\|}$.

Returning to the transformed formula for the sample correlation, $R = \dfrac{\dot{\mathrm{X}}^T\dot{\mathrm{Y}}}{\|\dot{\mathrm{X}}\|\|\dot{\mathrm{Y}}\|}$, we rewrite it as $R = cos\,\theta$, where θ is the angle between the centered vectors $\dot{\mathrm{X}}$ and $\dot{\mathrm{Y}}$ in $\mathbb{R}^n$. This is cool enough already, but the real coolness is that we've just proven one of the main properties of the sample correlation, that $-1 \le R \le 1$. Normally, the proof would require a fair amount of ordinary algebra, and doesn't provide the geometrical insight acquired from the linear algebraic proof above.

Special Cases: $R = 0$ and $R = \pm 1$

Two special cases of the sample correlation are when $R = 0$ and when $R = \pm 1$. Although they are almost never encountered in practice, they help to visualize the geometry of correlation in simple regression.

Special Case: $R = 0$

From the transformed formula $R = \dfrac{\dot{\mathrm{X}}^T \dot{\mathrm{Y}}}{\|\dot{\mathrm{X}}\| \|\dot{\mathrm{Y}}\|}$, $R = 0$ if and only if $\dot{\mathrm{X}}^T \dot{\mathrm{Y}} = 0$ if and only if $\dot{\mathrm{Y}} \perp \dot{\mathrm{X}}$.

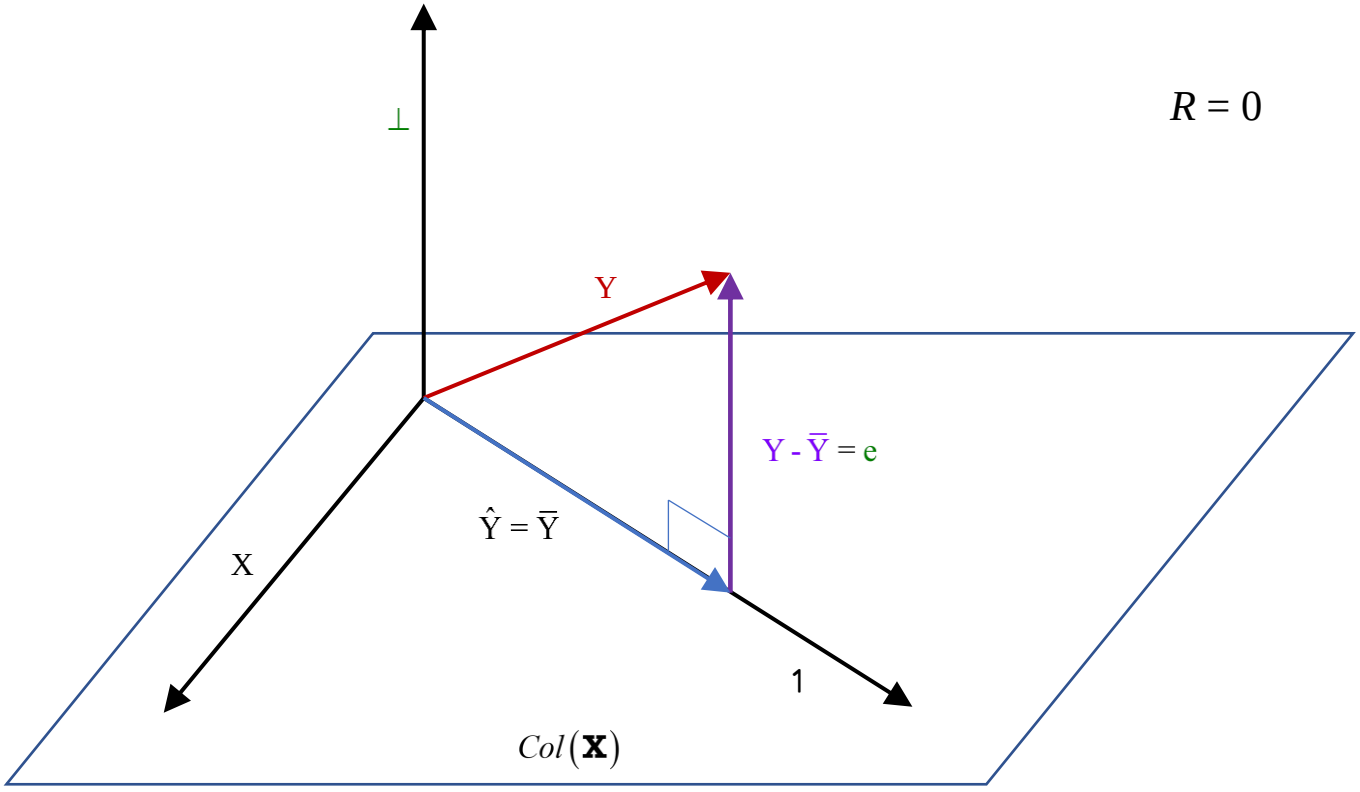
The following are always true in regression:

- $\dot{\mathrm{Y}} \perp 1$, where 1 is the n-vector of ones, because $\dot{\mathrm{Y}}^T 1 = \left(\mathrm{Y} - \bar{\mathrm{Y}}\right)^T 1 = \sum \left(Y_i - \bar{Y}\right) = 0$
- Similarly, $\dot{\mathrm{X}} \perp 1$ because $\sum \left(X_i - \bar{X}\right) = 0$

In addition, when the sample correlation is zero, $\dot{\mathrm{Y}} \perp \dot{\mathrm{X}}$.

The vectors 1 and $\dot{\mathrm{X}}$ are in $Col(\mathbf{X})$, and are independent because they are orthogonal. Furthermore, both are orthogonal to $\dot{\mathrm{Y}}$ when $R = 0$, so $\dot{\mathrm{Y}}$ is orthogonal to the column space of the design matrix, $Col(\mathbf{X})$, when $R = 0$. This immediately implies all of the following. (See sketch below)

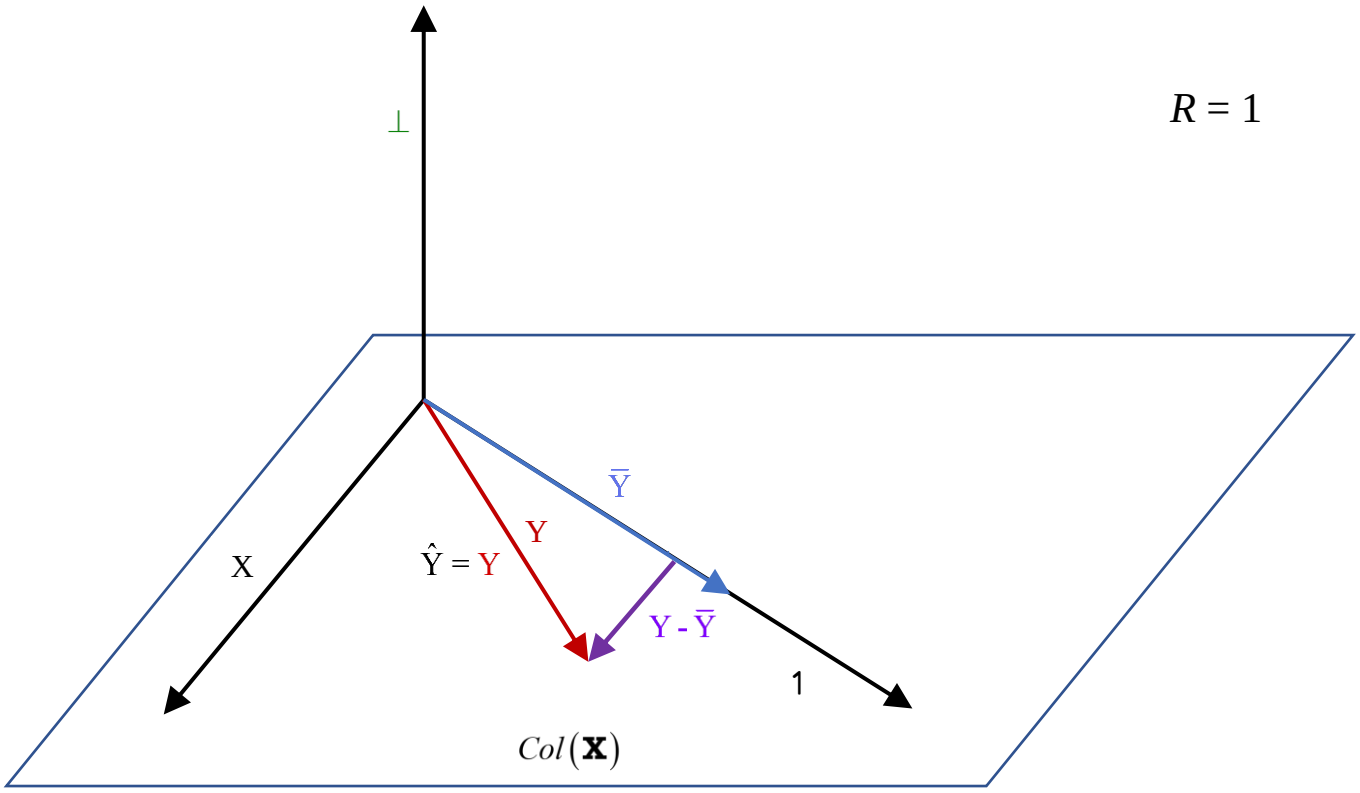
- $Y - \bar{Y} = e$, i.e., the residuals are just the deviations, $e_i = Y_i - \bar{Y}$.
- $\hat{Y} = \bar{Y}$, i.e., the fitted values are just the average value of the response for the sample because $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{Y}$ when $R = 0$. ($\hat{Y}_i = \bar{Y}$ for all X_i, and the fitted regression line is horizontal.)
- The vector of observations Y lies directly above the vector of ones, 1, because the vector X plays no role in fitting values when $\hat{\beta}_1 = 0$.



Special Case: $R = \pm 1$

From $R = cos\,\theta$, $R = \pm 1$ implies that $\dot{Y} \,||\, \dot{X}$. Since $\bar{Y} = \bar{y}1$ always lies in $Col(\mathbf{X})$, if $\dot{Y} = (Y - \bar{Y})$ also lies in $Col(\mathbf{X})$, then so must Y. This makes sense when we reflect that all points in the scatterplot line up along the regression line when $R = \pm 1$, so the residuals are all 0. Finally, when $R = \pm 1$, the fitted values are just the observations themselves, i.e., $\hat{Y} = Y$. Contrast this with the situation when $R = 0$ and $\hat{Y} = \bar{Y}$.

Because $\dot{Y} \,||\, \dot{X}$ and Y lies in $Col(\mathbf{X})$ when $R = \pm 1$, it's tempting to assume that Y is parallel to X in this case, but it need not be. However, if the regression line passes through the origin and $R = \pm 1$, then $Y = \hat{\beta}_1 X$ and the vector of observations Y is parallel the vector X.



Centering and Regression

Centering can also be used to obtain the estimated regression coefficients. There are several ways to accomplish this depending upon your goals, but we'll look at one simple approach.

Consider the following simple regression model: $\dot{\boldsymbol{Y}} = \hat{\beta}_1 \dot{X} + \boldsymbol{\varepsilon}$. So, what happened to the intercept term $\hat{\beta}_0$? Well, one effect of centering is that the fitted regression line will always pass through the origin, so we don't need an intercept term in the model. (But have no fear, because $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$ as before.)

The Design Matrix in Simple Regression Using Centering

Centering results in a different design matrix $\dot{\mathbf{X}}$ because the first column of ones is absent (because there is no intercept term in the model) and the remaining columns involve a centered vector for each predictor variable. In simple regression there is only one predictor, so the design matrix is particularly simple, it is the simply the vector $\dot{\mathbf{X}} = \dot{\mathbf{X}} = \begin{bmatrix} \dot{X}_1 \\ \dot{X}_2 \\ \vdots \\ \dot{X}_n \end{bmatrix}$. The equations leading to the least-squares estimators of the regression coefficients remains the same minus the missing $\hat{\beta}_0$, $\hat{\beta} = \left(\dot{\mathbf{X}}^T \dot{\mathbf{X}}\right)^{-1} \dot{\mathbf{X}}^T \dot{\mathrm{Y}}$.

In simple regression only $\hat{\beta}_1$ needs to be estimated this way, to the equation becomes $\hat{\beta}_1 = \left(\dot{\mathrm{X}}^T \dot{\mathrm{X}}\right)^{-1} \dot{\mathrm{X}}^T \dot{\mathrm{Y}}$, but $\dot{\mathrm{X}}^T \dot{\mathrm{X}}$ is just a number, so this becomes $\hat{\beta}_1 = \dfrac{\dot{\mathrm{X}}^T \dot{\mathrm{Y}}}{\dot{\mathrm{X}}^T \dot{\mathrm{X}}}$.

If this looks familiar to those who've had vector calculus, this states that the vector of fitted values $\hat{\mathrm{Y}}$ is the vector projection of $\dot{\mathrm{Y}}$ onto $\dot{\mathrm{X}}$, i.e., $\hat{\mathrm{Y}} = \mathbf{X}\hat{\beta}_1 = \dot{\mathrm{X}} \dfrac{\dot{\mathrm{X}}^T \dot{\mathrm{Y}}}{\dot{\mathrm{X}}^T \dot{\mathrm{X}}} = \left(\dfrac{\dot{\mathrm{X}}^T \dot{\mathrm{Y}}}{\dot{\mathrm{X}}^T \dot{\mathrm{X}}} \right) \dot{\mathrm{X}} = proj_{\dot{\mathrm{X}}} \dot{\mathrm{Y}}$. How cool is that?

Those of you who suffered through Math 32 with professor Hauser now understand why I put vector projections on every vector calculus exam.

Example

Using our evergreen example:

x	1	2	4	5
y	8	4	6	2

We create the centered vectors $\dot{\mathrm{X}} = \begin{bmatrix} -2 \\ -1 \\ 1 \\ 2 \end{bmatrix}$, $\dot{\mathrm{Y}} = \begin{bmatrix} 3 \\ -1 \\ 1 \\ -3 \end{bmatrix}$ (**Note:** A centered vectors is just the vector of deviations.)

Then, $\hat{\beta}_1 = \dfrac{\dot{\mathrm{X}}^T\dot{\mathrm{Y}}}{\dot{\mathrm{X}}^T\dot{\mathrm{X}}} = \dfrac{\begin{bmatrix}-2 & -1 & 1 & 2\end{bmatrix}\begin{bmatrix}3\\-1\\1\\-3\end{bmatrix}}{\begin{bmatrix}-2 & -1 & 1 & 2\end{bmatrix}\begin{bmatrix}-2\\-1\\1\\2\end{bmatrix}} = \dfrac{-10}{10} = -1$.

Also, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} = 5 - (-1)(3) = 8$.

In Statgraphics, I've entered the centered data below.

XC	YC
-2	3
-1	-1
1	1
2	-3

Regressing YC on XC produces the following *Analysis* window,

Coefficients

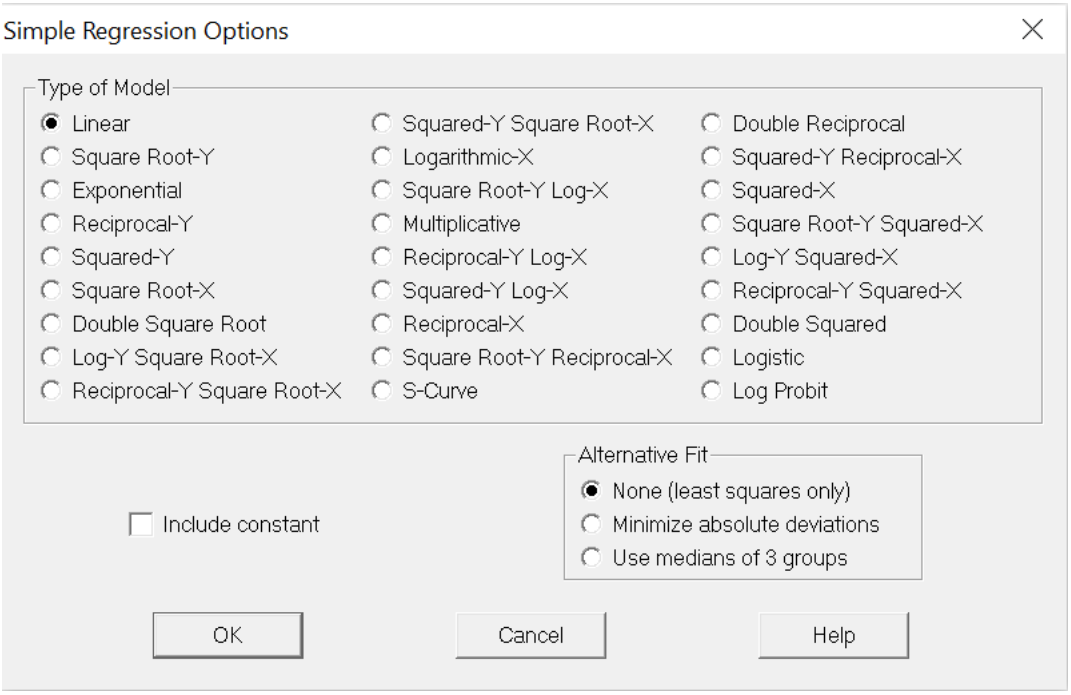
	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Intercept	0	1.11803	0	1.0000
Slope	-1.0	0.707107	-1.41421	0.2929

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	10.0	1	10.0	2.00	0.2929
Residual	10.0	2	5.0		
Total (Corr.)	20.0	3			

Correlation Coefficient = -0.707107
R-squared = 50.0 percent
R-squared (adjusted for d.f.) = 25.0 percent
Standard Error of Est. = 2.23607

Notice that the estimated intercept is zero, as expected. Alternatively, we could have unchecked the Include Constant box in the Analysis Options window.



Below is the output from Statgraphics. Notice that the estimated slope and correlation are unchanged, but the Mean Square Error is different because we are working with a different model. In this case, the *MSE* from the previous output is the one we want, as can be seen by rerunning the regression one more time on the original, uncentered, data.

Coefficients

	Least Squares	*Standard*	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
Slope	-1.0	0.57735	-1.73205	0.1817

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	10.0	1	10.0	3.00	0.1817
Residual	10.0	3	3.33333		
Total	20.0	4			

Correlation Coefficient = -0.707107
R-squared = 50.0 percent
R-squared (adjusted for d.f.) = 50.0 percent
Standard Error of Est. = 1.82574