Today’s lab will begin with a demo of key functions for working with projections. Fill in functions as we go and then, use the remaining time with a partner to replicate models from Lecture 1 as directed below.

**Putting lines through Pearson’s dataset**

```r
require(UsingR)
data(father.son)
summary(father.son)

fheight       sheight
Min. :59.01  Min. :58.51
1st Qu.:65.79 1st Qu.:66.93
Median :67.77 Median :68.62
Mean :67.69 Mean :68.68
3rd Qu.:69.60 3rd Qu.:70.47
Max. :75.43 Max. :78.36

par(mfrow=c(1,1))
plot(father.son, main="Pearson's height data",xlab="Father",ylab="Son", pch=20)
```

1. **Center the data. Produce a scatter plot of the data.**

For a given slope \( m \), the projection of a point \((x, y)\) onto the line with slope \( m \) through the origin can be computed as

\[
P(x, y) = \frac{(x, y) \cdot (1, m)}{1 + m^2}(1, m).
\]

That is the point \((x, y)\) gets new coordinates on the line with slope \( m \) through the origin by just scaling the point \((1, m)\) by a certain amount.
2. Write a function that takes a point \((x, y)\) and a line with slope \(m\) and returns the projection of the point, as well as the residual (the point minus its projection).

3. For Pearson’s centered data, verify that the sum of the squared lengths of the projections plus the sum of the squared lengths of the residuals is equal to the sum of the squared lengths of the centered data.

4. Write a function that takes a slope \(m\) and computes the sum of squared lengths of the projections. Plot it and find the slope that has the largest sum of squared lengths of the projections. What is the slope of the regression line of father onto son. Does it match? Should it?

5. Write a function that takes a slope \(m\) and computes the sum of squared lengths of the residuals. Plot it and find the slope that has the smallest sum of squared lengths of the residuals. Is it the same as one of the slopes in 4. Which one?

6. Given the slope from 4, create a vector \(v = (1, \text{slope\_from\_4})\). Now, with \(M = \text{cov}(\text{father}, \text{son})\) compute \(M \%*% v\). Compare this vector to \(v\).

Models from Lecture 1

We consider two models for data in Lecture 1. The first was

\[ X \sim \text{rnorm}; Y = 2X + \text{rnorm}. \]

1. Generate a large sample from this model. Make a scatter plot. What is the slope of the regression line of \(Y\) onto \(X\)?

2. Use your function above to find the line with the largest squared lengths of the residuals for this data. Does it have the same as the slope as the regression line?

The other model was

\[ Z \sim \text{rnorm}; X = Z + \cdot \text{rnorm}; Y = 2 \cdot Z + \sigma \cdot \text{rnorm}. \]

3. Generate a large sample from this model. Make a scatter plot. What is the slope of the regression line of \(Y\) onto \(X\)?

4. Use your function above to find the line with the largest squared lengths of the residuals for this data. Does it have the same as the slope as the regression line?