Visualizing Simultaneous Linear Equations, Geometric Vectors, and Least-Squares Regression with the **matlib** Package for R.

John Fox [1] Michael Friendly [2]

[1]McMaster University
Hamilton, Ontario, Canada

[2]York University
Toronto, Ontario, Canada

June 2016

Introduction

- Since the 1970s linear algebra and associated vector geometry have featured prominently in our teaching of applied statistics to social science and behavioural science graduate students.
- Many ideas, particularly in linear models and multivariate methods are clarified by the underlying algebra and geometry.
- We have used a variety of software to teach these ideas using tools of our own devising written in APL, SAS/IML, and, most recently, R.
- We've recently collected, formalized, and extended these tools in the **matlib** package for R.
- This talk focuses on the 2D and 3D visualization features of the **matlib** package — for visualizing linear simultaneous equations and geometric vectors, and the application of the latter to visualizing linear least-squares regression.

Visualizing Solutions to Linear Simultaneous Equations

- Representing linear equations as lines or planes makes it much easier for students to understand the properties of solutions to systems of linear equations and to link these solutions to ideas in linear algebra.
- This is particularly true in the 3D case.
- Some examples:

Visualizing Solutions to Linear Simultaneous Equations

An Over-determined Equation System in 2 Unknowns

```
> A <- matrix(scan(), byrow=TRUE, nrow=3)
1: 2 -2
3: 1 -1
5: 4  4
7:

> b <- 1:3
> showEqn(A, b, simplify=TRUE)
2*x1 - 2*x2  =  1
  x1   - x2  =  2
4*x1 + 4*x2  =  3
```

Visualizing Solutions to Linear Simultaneous Equations

An Over-determined Equation System in 2 Unknowns

```
> gaussianElimination(A, b, verbose=TRUE, fractions=TRUE)

Initial matrix:
     [,1] [,2] [,3]
[1,]  2   -2    1
[2,]  1   -1    2
[3,]  4    4    3

row: 1

 exchange rows 1 and 3
     [,1] [,2] [,3]
[1,]  4    4    3
[2,]  1   -1    2
[3,]  2   -2    1
```


Visualizing Solutions to Linear Simultaneous Equations

An Over-determined Equation System in 2 Unknowns

```
. . .

 multiply row 2 by 2 and add to row 3
     [,1] [,2] [,3]
[1,]    1    0  5/8
[2,]    0    1  1/8
[3,]    0    0  3/2
```

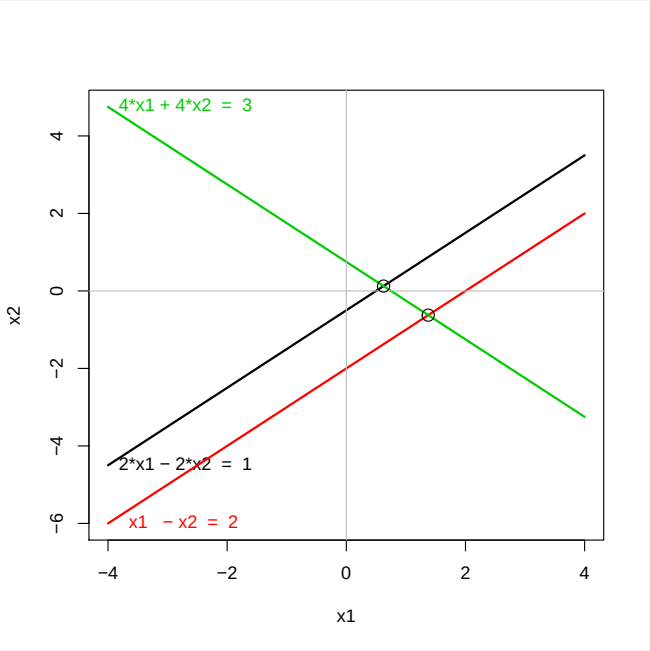
That is, $x_1 = 5/8, x_2 = 1/8, 0 = 3/2$.

Visualizing Solutions to Linear Simultaneous Equations

An Over-determined Equation System in 2 Unknowns

plotEqn(A, b)

- The source of the inconsistency in the equations is clearly revealed in the graph.



Visualizing Solutions to Linear Simultaneous Equations

An Equation System in 3 Unknowns with a Unique Solution

```
> A1 <- matrix(scan(), byrow=TRUE, nrow=3)
1: 2 -2  0
4: 1 -1  1
7: 4  4 -4

> b1 <- 1:3

> showEqn(A1, b1)
2*x1 - 2*x2 + 0*x3  =  1
1*x1 - 1*x2 + 1*x3  =  2
4*x1 + 4*x2 - 4*x3  =  3
```

Visualizing Solutions to Linear Simultaneous Equations

An Equation System in 3 Unknowns with a Unique Solution

```
> gaussianElimination(A1, B1=b, verbose=TRUE, fractions=TRUE)

Initial matrix:
     [,1] [,2] [,3] [,4]
[1,]  2   -2    0    1
[2,]  1   -1    1    2
[3,]  4    4   -4    3

row: 1

 exchange rows 1 and 3
     [,1] [,2] [,3] [,4]
[1,]  4    4   -4    3
[2,]  1   -1    1    2
[3,]  2   -2    0    1
```


Visualizing Solutions to Linear Simultaneous Equations

An Equation System in 3 Unknowns with a Unique Solution

```
. . .

 multiply row 3 by 1/2 and add to row 2
     [,1] [,2] [,3] [,4]
[1,]    1    0    0 11/8
[2,]    0    1    0  7/8
[3,]    0    0    1  3/2
```

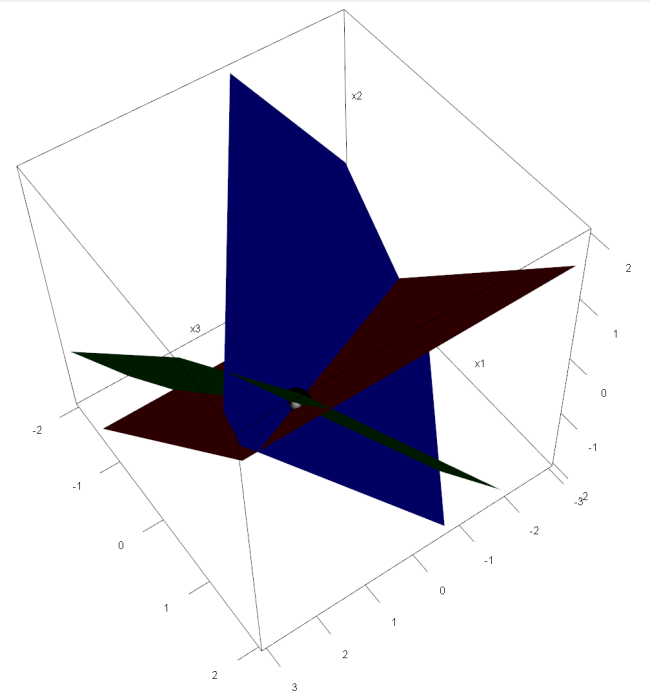
Thus, $x_1 = 11/8, x_2 = 7/8, x_3 = 3/2$.

Visualizing Solutions to Linear Simultaneous Equations

An Equation System in 3 Unknowns with a Unique Solution

```
plotEqn3d(A1, b1)
```

- The 3 planes intersect at a point, determining the solution to the system of equations.
- The original **rgl** graph is manipulable.



Visualizing Solutions to Linear Simultaneous Equations

An Under-determined Equation System in 3 Unknowns

```
> A2 <- matrix(scan(), byrow=TRUE, nrow=3)
1: 2 -2   0
4: 1 -1   1
7: 1 -1  -1

> b2 <- c(1, 2, -1)

> gaussianElimination(A2, B=b2, fractions=TRUE)
     [,1] [,2] [,3] [,4]
[1,]    1   -1    0  1/2
[2,]    0    0    1  3/2
[3,]    0    0    0    0
```

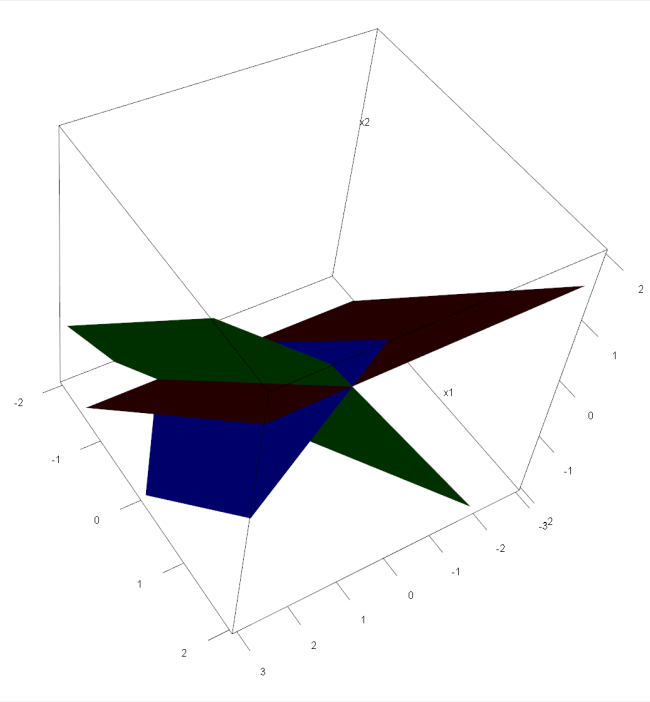
That is, $x_1 - x_2 = 1/2, x_3 = 3/2$.

Visualizing Solutions to Linear Simultaneous Equations

An Under-determined Equation System in 3 unknowns

```
plotEqn3d(A2, b2)
```

- The three planes intersect in a common line, representing the infinity of solutions $x_2 = x_1 - 1/2, x_3 = 3/2$.



Visualizing Solutions to Linear Simultaneous Equations

An Over-determined Equation System in 3 Unknowns

```
> A3 <- A2
> b3 <- c(2, 2, 2)
> gaussianElimination(A3, B=b3, fractions=TRUE)
     [,1] [,2] [,3] [,4]
[1,]    1   -1    0    1
[2,]    0    0    1    1
[3,]    0    0    0    2
```

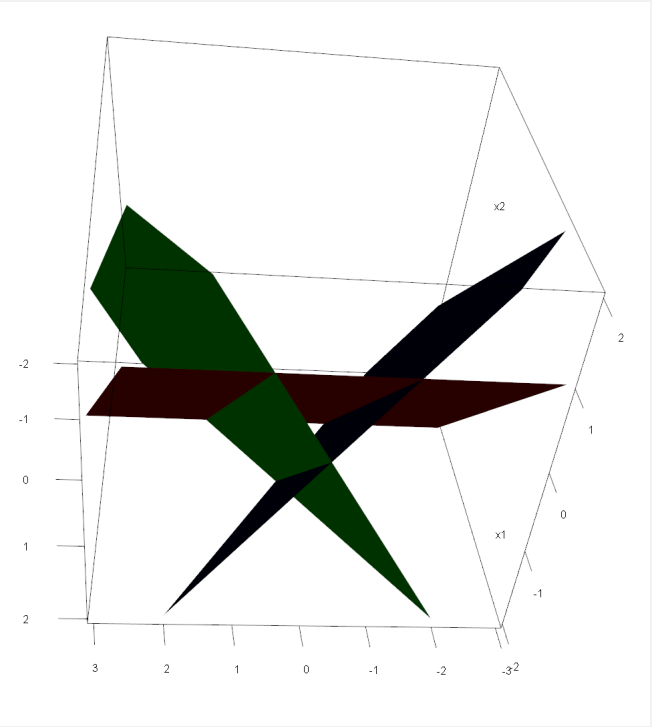
That is, $x_1 - x_2 = 1, x_3 = 1, 0 = 2$.

Visualizing Solutions to Linear Simultaneous Equations

An Under-determined Equation System in 3 Unknowns

```
plotEqn3d(A3, b3)
```

- The three planes fail to intersect, representing the absence of a solution.



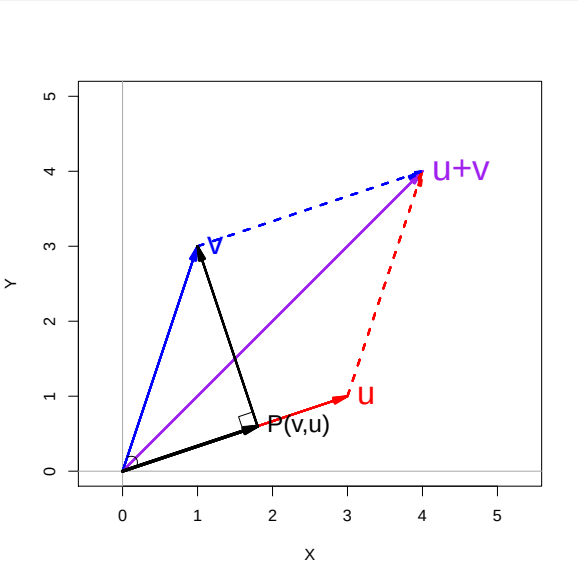
Visualizing Geometric Vectors

- Many ideas in linear algebra are illuminated by translating them into vector geometry.
- Most key ideas are expressible geometrically in 2 or 3 dimensions and can then be generalized algebraically to spaces of higher dimension.
- This is true as well of applications of vector geometry in statistics, in particular to linear models and to multivariate statistics.
- The **matlib** package incorporates facilities for drawing vectors in 2D and 3D space, the latter using the **rgl** package.

Visualizing Geometric Vectors in 2D

Illustrating the parallelogram rule of vector addition and $\perp$ projection:

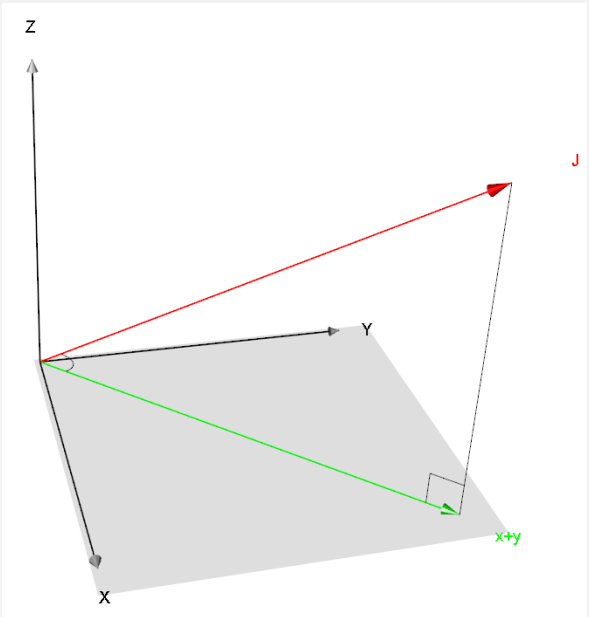
```
> u <- c(3,1); v <- c(1,3)
> plot(c(0, 5), c(0, 5), type="n",
+   xlab="X", ylab="Y", asp=1)
> abline(v=0, h=0, col="darkgray")
> vectors(rbind(u, v, "u+v"=u + v),
+   col=c("red", "blue", "purple"),
+   cex.lab=c(2, 2, 2.2))
> vectors(u + v, origin=u,
+   col="red", lty=2)
> vectors(u + v, origin=v,
+   col="blue", lty=2)
> vectors(Proj(v, u),
+   labels="P(v,u)", lwd=3)
> vectors(v, origin=Proj(v, u))
> corner(c(0, 0), Proj(v, u), v, d=0.2)
> arc(v, c(0, 0), u, d=0.2)
```



Visualizing Geometric Vectors in 3D

Orthogonal projection of j on $\{x, y\}$:

```
> open3d()
> E <- diag(3)
> rownames(E) <- c("x", "y", "z")
> vectors3d(E, lwd=2)
> vectors3d(c(1, 1, 1),
+   labels=c("","j"), color="red",
+   lwd=2)
> vectors3d(c(1, 1, 0),
+   labels=c("", "x+y"),
+   color="green", lwd=2)
> planes3d(0, 0, 1, 0, col="gray",
+   alpha=0.2)
> segments3d(rbind(c(1, 1, 1),
+   c(1, 1, 0)))
> arc(c(1, 1, 1), c(0, 0, 0),
+   c(1, 1, 0))
> corner(c(0, 0, 0), c(1, 1, 0), c(1, 1, 1))
```



Visualizing the Vector Geometry of Regression

- A particularly compelling application of geometric vectors is to least-squares linear regression.
- Many ideas in regression are clarified and rendered intuitive by vector representation, including (see, e.g., Fox, 2016, , Chap. 10):
  - least-squares fit
  - the ANOVA decomposition of the total sum of squares into components
  - degrees of freedom
  - unbiased estimation of the error variance
  - Simpson's paradox
  - more generally, the distinction between partial and marginal relationships
  - collinearity
- We develop some of these ideas below.

Visualizing the Vector Geometry of Regression

- The `regvec3d` function in the **matlib** package creates a geometric vector representation of a least-squares regression in 3D vector space.
- This is most straightforward when there are 2 predictors:
  - Treating the predictors x_1 and x_2 as fixed, and writing them as deviations from their means, $\mathbf{x}_1 = \{x_{1i} - \overline{x}_1\}$ and $\mathbf{x}_2 = \{x_{2i} - \overline{x}_2\}$, and the response y as deviations from its unconditional expectation, $\mathbf{y} = \{y_i - E(y)\}$, we eliminate the constant in the regression model, obtaining
$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\varepsilon}$$
where $\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right)$.
  - The vectors $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{y}$ span a subspace of dimension 3 of the n-dimensional observation space, and other vectors of interest (e.g., $\boldsymbol{\varepsilon}$) lie in this subspace.
- When there are more than 2 predictors, we can pick 2 and condition on the remaining predictors by residualizing.

Visualizing the Vector Geometry of Regression

- The estimated regression model is

$$\mathbf{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \mathbf{e}$$

  where the residual vector $\mathbf{e} = \{e_i\}$ lies in the 3D $\{\mathbf{x_1}, \mathbf{x_2}, \mathbf{y}\}$ space, and the vector of fitted values $\widehat{\mathbf{y}} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2$ lies in the 2D regressor subspace spanned by $\mathbf{x}_1$ and $\mathbf{x}_2$.
- Geometrically, minimizing $\sum e_i^2$ is equivalent to minimizing the length $\|\mathbf{e}\|$ of the residual vector, which implies that $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the $\{\mathbf{x}_1, \mathbf{x}_2\}$ subspace — a derivation of the LS fit without calculus.
- The ANOVA for the regression follows from the Pythagorean Theorem as

$$\|\mathbf{y}\|^2 = \|\widehat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$$

- and the correlation between x and y is simply the cosine of the angle between their vectors.

Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

- We illustrate with O. D. Duncan's regression of the prestige of 45 U.S. occupations in 1950 on the education and income levels of the occupations (Duncan, 1961), with data in the `Duncan` data frame in the **car** package (Fox and Weisberg, 2011).
- Variables:

`prestige`	Percentage ratings of good or better in a national survey.
`education`	Percentage of high school graduates in the 1950 Census.
`income`	Percentage earning $3500 or more in the 1950 Census.

Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

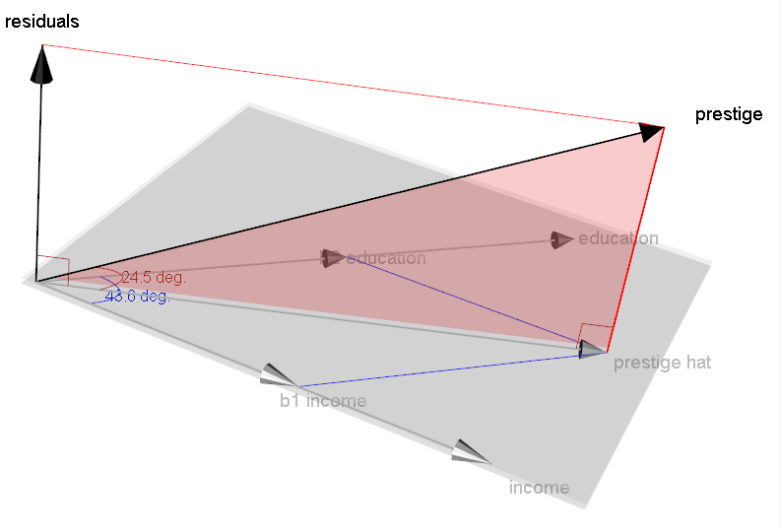
```
> summary(lm(prestige ~ income + education, data=Duncan))

. . .

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05
education    0.54583    0.09825   5.555 1.73e-06

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282, Adjusted R-squared:   0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```


Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

A basic vector visualization of Duncan's regression:

```
> duncan.vecs <- regvec3d(prestige ~ income + education,
+   data=Duncan)
> plot(duncan.vecs)
```

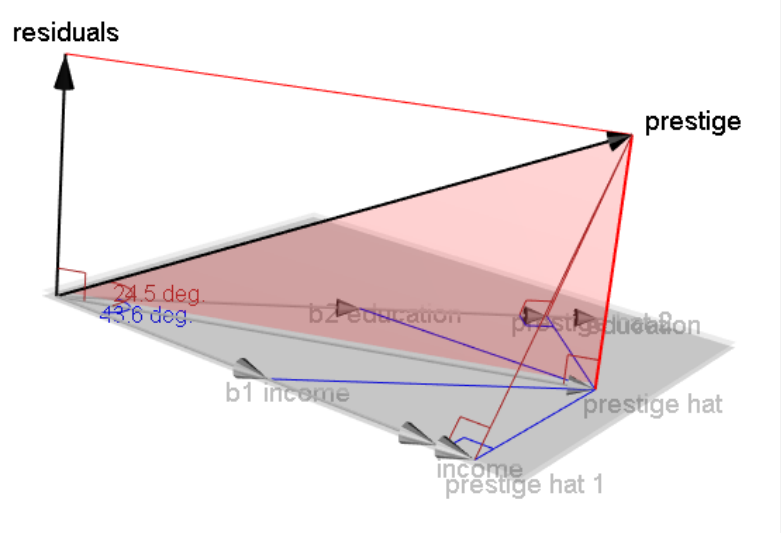


Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

Adding the marginal regressions:

```
> plot(duncan.vecs, show.marginal=TRUE)
```

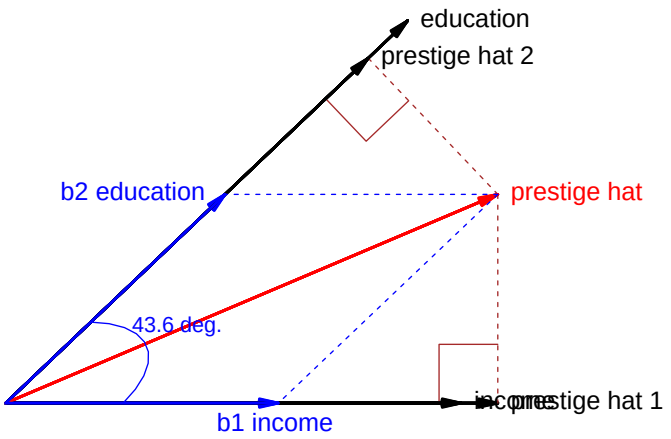


Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

The distinction between the partial and marginal regression coefficients is clarified by examining the 2D regressor subspace:

```
> plot(duncan.vecs,
+   dimension=2,
+   show.marginal=TRUE)
```

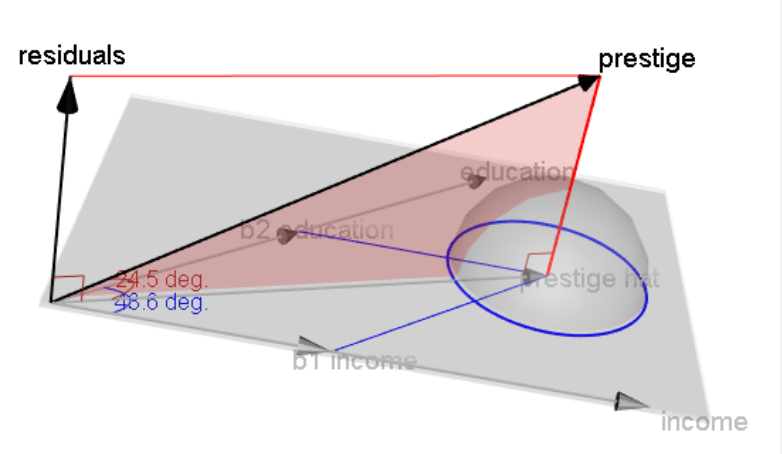


Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

Finally, we add a representation of sampling variability by including the 3D projection of the error hypersphere, scaled so that its projections on the **x** axes show confidence intervals for the standardized regression coefficients:

```
> duncan.vecs.2 <- regvec3d(prestige ~ income + education,
+   data=Duncan, scale=TRUE)
> plot(duncan.vecs.2, error.sphere="y.hat")
```

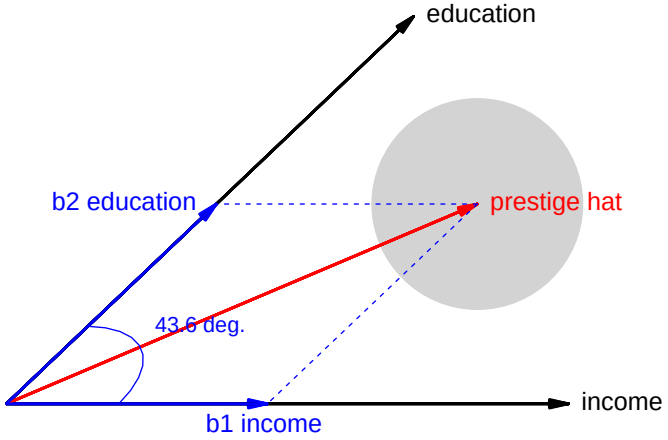


Visualizing the Vector Geometry of Regression

Duncan's Occupational Prestige Regression

... or, in the regressor plane:

```
> plot(duncan.vecs.2,
+   dimension=2,
+   error.sphere="y.hat")
```



Summary and Conclusions

- **matlib** began an an attempt to reproduce in R software tools we had used in teaching linear algebra and vector geometry over many years.
- It has developed into a package for two audiences:
  - those who would like to explore or explain the inner workings of methods of linear algebra
  - those who would like to illustrate statistical ideas geometrically, particularly for linear models
- 2D vector diagrams are relatively straightforward; 3D diagrams with **rgl** are more of a challenge, but can prove particularly rewarding.
- We intend to continue to develop this geometric approach to explicating statistical ideas in R.

References

Duncan, O. D. (1961). A socioeconomic index for all occupations. In Reiss, Jr., A. J., editor, *Occupations and Social Status*, pages 109–138. Free Press, New York.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. Thousand Oaks CA, 3rd edition.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks CA, 2nd edition.