The Geometry of Linear Models

STATS 762 – Lecture Notes 2

Arden Miller

Department of Statistics, University of Auckland

Where Now?

In the first set of lecture notes, we described the linear model and its estimation from a geometric viewpoint.

In this set of lecture slides we will start to explore doing statistical inference based on the estimated linear model.

We need to review some basic matrix algebra.

Some Useful Matrix Algebra

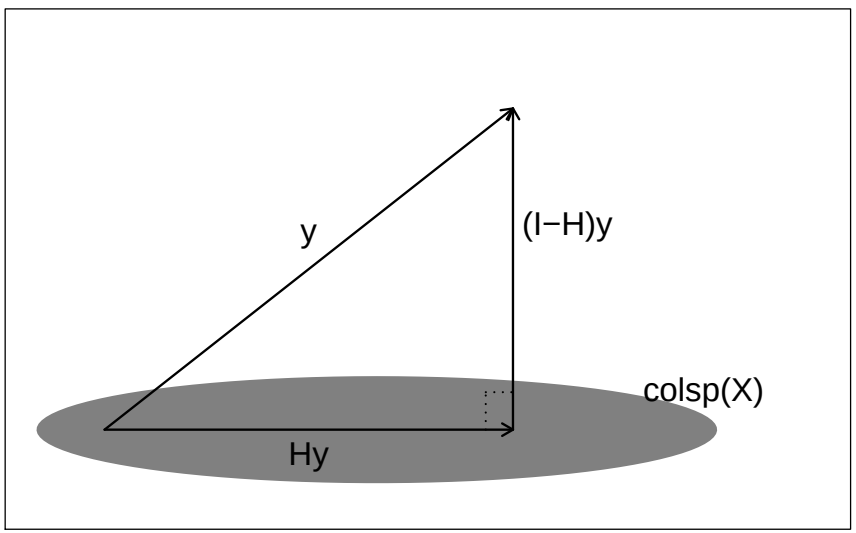
For matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ with compatible dimensions:

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
- $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$
- $(c + d)\mathbf{A} = c\mathbf{A} + d\mathbf{A}$
- $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$
- $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$ if $\mathbf{A}$ and $\mathbf{B}$ are both square
- $(\mathbf{A}^t)^{-1} = (\mathbf{A}^{-1})^t$
- $(c\mathbf{A})^t = c\mathbf{A}^t$
- $(\mathbf{A}^t)^t = \mathbf{A}$
- $(\mathbf{AB})^t = \mathbf{B}^t\mathbf{A}^t$
- $(c\mathbf{A} + d\mathbf{B})^t = c\mathbf{A}^t + d\mathbf{B}^t$

Fitting the Linear Model

Linear Model: $\mathbf{Y} = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\epsilon}$ where $\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}$.

Fitted Model: $\mathbf{Y} = \hat{\boldsymbol{\mu}}_{\mathbf{Y}} + \mathbf{r}$ where $\hat{\boldsymbol{\mu}}_{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.



$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\mathbf{Y}} &= \mathbf{H}\mathbf{Y} \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ \mathbf{r} &= (\mathbf{I} - \mathbf{H})\mathbf{Y} \end{aligned}$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$

Distributions of $\hat{\mu}_{\mathbf{Y}}$, $\hat{\beta}$ and $\mathbf{r}$

$\hat{\mu}_{\mathbf{Y}}$, $\hat{\beta}$ and $\mathbf{r}$ are all linear transformations of $\mathbf{Y}$ – i.e. they can be written as a matrix times $\mathbf{Y}$:

$$\hat{\mu}_{\mathbf{Y}} = \left[\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right]\mathbf{Y}$$

$$\hat{\beta} = \left[\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right]\mathbf{Y}$$

$$\mathbf{r} = [\mathbf{I} - \mathbf{H}]\,\mathbf{Y}$$

As a result we can derive each of their distributions from the distribution of $\mathbf{Y}$.

Linear Transformations of Random Vectors

The following properties of linear transformations of random vectors are extremely useful. Consider:

$$\mathbf{U} = \mathbf{M}\mathbf{V} + \mathbf{C}$$

where $\mathbf{M}$ and $\mathbf{C}$ are fixed and $\mathbf{V}$ is a random vector with mean vector $\boldsymbol{\mu}_{\mathbf{V}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{V}}$.

$$\underbrace{\begin{bmatrix} U_1 \\ \vdots \\ U_p \end{bmatrix}}_{\mathbf{U}} = \underbrace{\begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1k} \\ \vdots & \vdots & & \vdots \\ m_{p1} & m_{p2} & \cdots & m_{pk} \end{bmatrix}}_{\mathbf{M}} \underbrace{\begin{bmatrix} V_1 \\ \vdots \\ V_k \end{bmatrix}}_{\mathbf{V}} + \underbrace{\begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix}}_{\mathbf{C}}$$

Then $\boldsymbol{\mu}_{\mathbf{U}} = \mathbf{M}\boldsymbol{\mu}_{\mathbf{V}} + \mathbf{C}$ and $\boldsymbol{\Sigma}_{\mathbf{U}} = \mathbf{M}\boldsymbol{\Sigma}_{\mathbf{V}}\mathbf{M}^t$.

- Further, if $\mathbf{V}$ has a Normal distribution so does $\mathbf{U}$.

Distributions of $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$, $\hat{\boldsymbol{\beta}}$ and $\mathbf{r}$

Given the previous slide and

$$\mathbf{Y} \sim \mathrm{N}(\boldsymbol{\mu}_{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta},\ \boldsymbol{\Sigma}_{\mathbf{Y}} = \sigma^2\mathbf{I})$$

we can deduce the distributions of $\hat{\boldsymbol{\mu}}_{\mathbf{Y}}$, $\hat{\boldsymbol{\beta}}$ and $\mathbf{r}$:

$\hat{\boldsymbol{\mu}}_{\mathbf{Y}} \sim$

$\hat{\boldsymbol{\beta}} \sim$

$\mathbf{r} \sim$

Inference for $\hat{\boldsymbol{\beta}}$

We will use $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta},\ \sigma^2 (\mathbf{X}^t\mathbf{X})^{-1})$ as the basis for inference about the regression coefficients.

- The diagonal elements of $\sigma^2 (\mathbf{X}^t\mathbf{X})^{-1}$ are the variances of the $\hat{\beta}_i$'s and the off-diagonal elements are covariances between pairs of $\hat{\beta}_i$'s. All of these depend on σ^2.
- If σ^2 is not known (which is almost always the case) then we need to estimate σ^2 before we can proceed.

Estimating σ^2

The standard way of estimating σ^2 is to divide the residual sum of squares by the residual degrees of freedom:

$$\hat{\sigma}^2 = \frac{\mathbf{r}^t\mathbf{r}}{n-k-1}$$

- We will justify the use of this estimator later. For now, we will just accept that it provides us with an unbiased estimate of σ^2

The Catheter Example: $\hat{\beta}$

For the catheter example we can calculate $\hat{\beta}$:

```
> x1<-c(42.8,63.5,37.5,39.5,45.5,38.5,
+       43.0,22.5,37.0,23.5,33.0,58.0)
> x2<-c(40.0,93.5,35.5,30.0,52.0,17.0,
+       38.5, 8.5,33.0, 9.5,21.0,79.0)
> X<-cbind(1,x1,x2)
> y<-matrix(c(37,50,34,36,43,28,37,20,34,30,38,47),12,1)
> BETAhat<- solve(t(X)%*%X)%*%t(X)%*%y
> BETAhat
          [,1]
   20.3757645
x1  0.2107473
x2  0.1910949
```

The Catheter Example: $(\mathbf{X}^t\mathbf{X})^{-1}$

To get $(\mathbf{X}^t\mathbf{X})^{-1}$:

```
> XtXinv<-solve(t(X)%*%X)
> XtXinv
                            x1           x2
    4.9262358 -0.197172444  0.081695704
x1 -0.1971724  0.008363701 -0.003681904
x2  0.0816957 -0.003681904  0.001754749
```

- We need to multiply $(\mathbf{X}^t\mathbf{X})^{-1}$ by $\hat{\sigma}^2$ to get our estimated covariance matrix for $\hat{\boldsymbol{\beta}}$.

The Catheter Example: Residuals

For the catheter example we can calculate the residuals from:

```
> res<-(diag(12)-H)%*%y
> res
              [,1]
 [1,] -0.03954422
 [2,] -1.62559020
 [3,] -1.06265657
 [4,]  1.56687083
 [5,]  3.09829930
 [6,] -3.73814817
 [7,]  0.20494867
 [8,] -6.74188499
 [9,] -0.47954567
[10,]  2.85627283
[11,]  6.65658228
[12,] -0.69560408
```

The Catheter Example: $\hat{\sigma}^2$

To get $\hat{\sigma}^2$:

```
> sig2hat<-(t(res)%*%res)/(12-2-1)
> sig2hat
          [,1]
[1,] 14.27543
```

Thus to get the estimated covariance matrix for $\hat{\boldsymbol{\beta}}$

```
> as.numeric(sig2hat)*XtXinv
                          x1           x2
   70.324134 -2.81472140  1.16624129
x1 -2.814721  0.11939543 -0.05256076
x2  1.166241 -0.05256076  0.02504980
```

The Catheter Example: Inference for $\hat{\boldsymbol{\beta}}$

Given we now have $\hat{\boldsymbol{\beta}}$ and an estimated covariance matrix for $\hat{\boldsymbol{\beta}}$, we can do the standard types of statistical inference (find confidence intervals and do hypothesis tests.

- Since we had to estimate σ^2, we must use a t_{n-k-1} reference distribution.

To get a $(1-\alpha)100\%$ confidence interval for β_i:

$$\hat{\beta}_i \pm t_{n-k-1}(1-\alpha/2) \times \text{se}(\hat{\beta}_i)$$

To test H_0: $\beta_i = \text{constant}$, calculate

$$\text{t-stat} = \frac{\hat{\beta}_i - \text{constant}}{\text{se}(\hat{\beta}_i)}$$

$$\text{p-value} = 2 \times \Pr(t_{n-k-1} \geq |\text{t-stat}|)$$

The Catheter Example: Inference for $\hat{\beta}$

For example, to get a 95% confidence interval for β_1:

$$\begin{aligned} \hat{\beta}_1 &\pm t_9(.975) \times \text{se}(\hat{\beta}_1) \\ 0.211 &\pm 2.26\sqrt{.119} \\ 0.211 &\pm 0.782 \end{aligned}$$

Or to test H_0:$\beta_1 = 0$

$$\text{t-stat} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{0.211}{\sqrt{.119}} = 0.611$$

$$\text{p-value} = 2 \times \Pr(t_9 \geq |t - stat|) = 2 \times \Pr(t_9 \geq 0.611) = 0.56$$

The Catheter Example: Predicted Values

Suppose we wanted to predict the catheter length for a child that was 44 inches tall (X_1) and weighed 35 pounds (X_2).

We can proceed by applying the results from slide 6:

$$\text{estimate} = \begin{bmatrix} 1 & 44 & 35 \end{bmatrix} \begin{bmatrix} 20.376 \\ 0.211 \\ 0.191 \end{bmatrix} = 36.34$$

$$\text{est. var.} = \begin{bmatrix} 1 & 44 & 35 \end{bmatrix} \begin{bmatrix} 70.324 & -2.815 & 1.166 \\ -2.815 & 0.119 & -0.053 \\ 1.166 & -0.053 & 0.025 \end{bmatrix} \begin{bmatrix} 1 \\ 44 \\ 35 \end{bmatrix} = 4.21$$

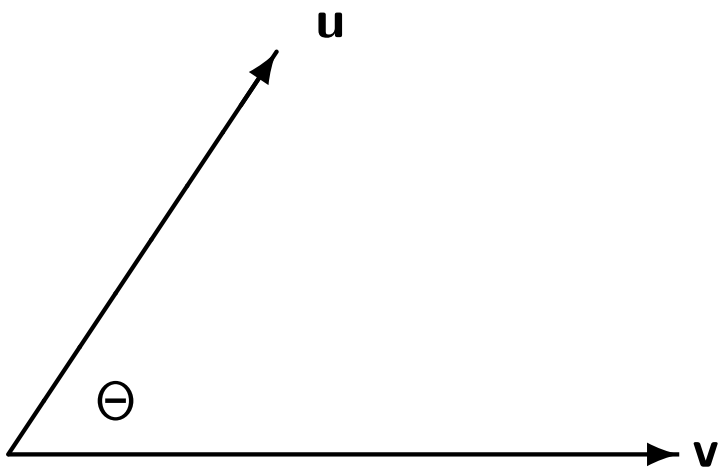
Orthogonality

Orthogonality will be a re-occurring concept in our discussion.

Therefore, we will take a bit of time now to elaborate on orthogonality as it relates to vectors and vector spaces.

The Angle between Two Vectors

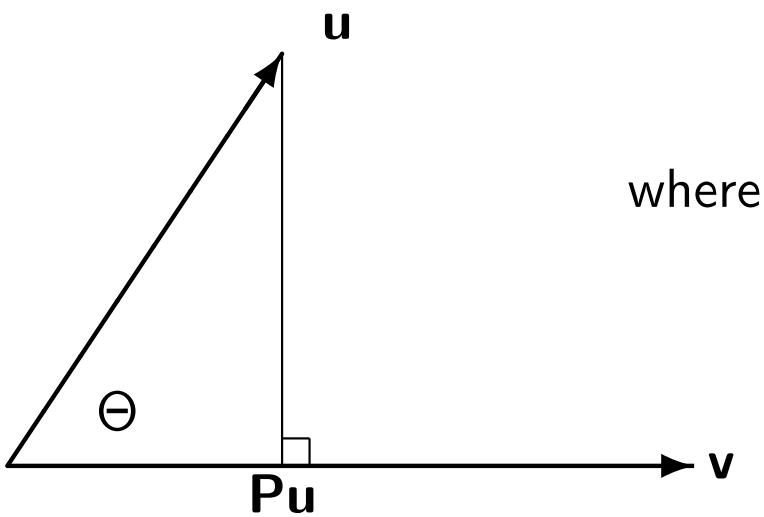
Suppose we want to find the angle between vectors $\mathbf{u}$ and $\mathbf{v}$:



$$\cos\Theta = \frac{\mathbf{u}^t\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

The Angle between Two Vectors

To derive this result, consider the projection of $\mathbf{u}$ onto $\mathbf{v}$:



where $\mathbf{P} = \mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t$

By definition: $\cos\Theta = \dfrac{\|\mathbf{P}\,\mathbf{u}\|}{\|\mathbf{u}\|}$

The Angle between Two Vectors

$$
\begin{aligned}
\|\mathbf{P}\,\mathbf{u}\|^2 &= (\mathbf{P}\mathbf{u})^t\mathbf{P}\mathbf{u} = \mathbf{u}^t\mathbf{P}^t\mathbf{P}\mathbf{u} \\
&= \mathbf{u}^t\left(\mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t\right)^t\mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t\mathbf{u} \\
&= \mathbf{u}^t\mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t\mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t\mathbf{u} \\
&= \mathbf{u}^t\mathbf{v}(\mathbf{v}^t\mathbf{v})^{-1}\mathbf{v}^t\mathbf{u} \\
&= \frac{\mathbf{u}^t\mathbf{v}\mathbf{v}^t\mathbf{u}}{\mathbf{v}^t\mathbf{v}} = \frac{(\mathbf{u}^t\mathbf{v})^2}{\|\mathbf{v}\|^2}
\end{aligned}
$$

Therefore $\|\mathbf{P}\,\mathbf{u}\| = \dfrac{\mathbf{u}^t\mathbf{v}}{\|\mathbf{v}\|}$ and thus $\cos\Theta = \dfrac{\|\mathbf{P}\,\mathbf{u}\|}{\|\mathbf{u}\|} = \dfrac{\mathbf{u}^t\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$

Orthogonal Things

Vectors **u** and **v** are orthogonal if they form a right angle which implies that $\cos\Theta = 0$ where Θ is the angle between **u** and **v**.

A vector **v** is orthogonal to a subspace S if it is orthogonal to every vector in S.

Subspaces S_1 and S_2 are orthogonal if every vector in S_1 is orthogonal to every vector in S_2.

Showing Things are Orthogonal Things

To show vectors $\mathbf{u}$ and $\mathbf{v}$ are orthogonal:

- Show $\mathbf{u}^t\mathbf{v} = 0$.

To show a vector $\mathbf{v}$ is orthogonal to a subspace S:

- Show $\mathbf{v}$ is orthogonal to each vector in a basis for S.
- Show $\mathbf{Pv} = \mathbf{0}$ where $\mathbf{P}$ is the orthogonal projection matrix for S.

To show subspaces S_1 and S_2 are orthogonal:

- Show that each vector in a basis for S_1 is orthogonal to each vector in a basis for S_2.
- Show that $\mathbf{P_1P_2} = \mathbf{0}$ where $\mathbf{P_1}$ and $\mathbf{P_2}$ are the projection matrices for S_1 and S_2.

The Orthogonal Complement of a Subspace of R^n

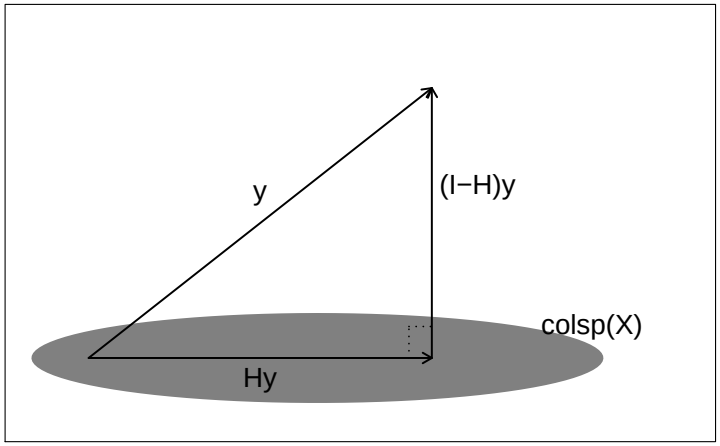
Let S be any subspace of R^n, then the set of all vectors that are orthogonal to S themselves form a subspace. This subspace is called the *orthogonal compliment* of S which we will denote as $S^{\perp}$.

- $\dim(S) + \dim(S^{\perp}) = n$.
- If we combine a basis for S with a basis for $S^{\perp}$, we get a basis for R^n.
- If $\mathbf{P}$ is the projection matrix for S, then $\mathbf{I} - \mathbf{P}$ is the projection matrix for $S^{\perp}$.
- For any vector $\mathbf{u} \in R^n$:

$$\|\mathbf{u}\|^2 = \|\mathbf{P}\mathbf{u}\|^2 + \|(\mathbf{I} - \mathbf{P})\mathbf{u}\|^2$$

Back to the Linear Model

Fitting the linear model can be thought of as projecting $\mathbf{y}$ on to the columnspace of $\mathbf{X}$.



The residual vector $\mathbf{r} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, is the component of $\mathbf{y}$ that is orthogonal to colsp($\mathbf{X}$). Therefore, it is an element of the orthogonal compliment of colsp($\mathbf{X}$) – we will call this the *error space*.

The Error Space

By definition, the error space contains all vectors that are orthogonal to colsp($\mathbf{X}$).

- The error space has dimension $n - k - 1$.
- The projection matrix on to the error space is $\mathbf{I} - \mathbf{H}$.
- The residual vector $\mathbf{r}$ is the orthogonal projection of $\mathbf{y}$ on to the error space.

Estimating σ^2

To lay the foundation for estimating σ^2, we will outline some relevant properties of a N($\mathbf{0}$, $\sigma^2\mathbf{I}$) vector – such as ϵ.

Let $\mathbf{V} = (V_1, \ldots V_n)^t$ be a N($\mathbf{0}$, $\sigma^2\mathbf{I}$) random vector.

- $V_1, \ldots V_n$ are independent, N(0, σ^2) random variables. Their joint density function is :

$$f(v_1, v_2, \ldots v_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(v_1^2+v_2^2+\ldots v_n^2)/2\sigma^2}$$

- Since $V_1/\sigma, \ldots V_n/\sigma$ are independent N(0, 1) random variables

$$\left(\sum_i V_i^2\right)/\sigma^2 \text{ has a } \chi^2_n \text{ distribution.}$$

Estimating σ^2

Translating the results from the previous slide to the random vector $\mathbf{V}$:

- The density function $f(\mathbf{v})$ of $\mathbf{V}$ is a function of $\|\mathbf{v}\|$:

$$f(\mathbf{v}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\|\mathbf{v}\|^2/2\sigma^2}$$

  - $f(\mathbf{v})$ is perfectly symmetric about the origin and $f(\mathbf{v})$ decreases as $\|\mathbf{v}\|$ increases.
- $\|\mathbf{V}\|^2$ is the sum of its squared elements:

$$\|\mathbf{V}\|^2 = \mathbf{V}^t\mathbf{V} = \sum_i V_i^2$$

and thus $\|\mathbf{V}\|^2/\sigma^2$ has a χ^2_n distribution.

The Distribution of $\|\mathbf{V}\|^2$

Given that $\|\mathbf{V}\|^2/\sigma^2 \sim \chi^2_n$ and $\mathrm{E}(\chi^2_n) = n$:

$$\mathrm{E}\left(\frac{\|\mathbf{V}\|^2}{\sigma^2}\right) = n \quad \rightarrow \quad \mathrm{E}\left(\frac{\|\mathbf{V}\|^2}{n}\right) = \sigma^2$$

- Thus for an observed response vector $\mathbf{v}$:

  $\|\mathbf{v}\|^2/n$ is an unbiased estimate of σ^2.

The Distributions of Projections

Now, consider a projection matrix $\mathbf{P}$ which projects $\mathbf{V} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$ on to a p-dimensional subspace of R^n.

- Due the symmetrical nature of the density $f(\mathbf{v})$, the distribution of $\mathbf{PV}$ only depends on the dimension of the subspace.
- The projection $\mathbf{PV}$ has a p-dimensional multivariate Normal distribution.
- The squared length of the projection $\|\mathbf{PV}\|^2/\sigma^2$ has a χ^2_p distribution.

Estimating σ^2

The residual vector $\mathbf{r}$ is the orthogonal projection of $\mathbf{y}$ on to the error space. This is equivalent to projecting $\boldsymbol{\epsilon}$ onto the error space (why?).

As a consequence, $\|\mathbf{r}\|^2/\sigma^2$ has a χ^2_{n-k-1} distribution which justifies using

$$\hat{\sigma}^2 = \frac{\mathbf{r}^t\mathbf{r}}{n-k-1}$$

as an estimate of σ^2.